# Polygonal Unadjusted Langevin Algorithms

# Polygonal Unadjusted Langevin Algorithms: Creating stable and efficient adaptive algorithms for neural networks *

Dong-Young Lim [1] and Sotirios Sabanis[1, 2]

[1]School of Mathematics, The University of Edinburgh, UK.
[2]The Alan Turing Institute, UK.

May 31, 2021

**Abstract**

We present a new class of adaptive stochastic optimization algorithms, which overcomes many of the known shortcomings of popular adaptive optimizers that are currently used for the fine tuning of artificial neural networks (ANNs). Its underpinning theory relies on advances of Euler's polygonal approximations for stochastic differential equations (SDEs) with monotone coefficients. As a result, it inherits the stability properties of tamed algorithms, while it addresses other known issues, e.g. vanishing gradients in ANNs. In particular, we provide an nonasymptotic analysis and full theoretical guarantees for the convergence properties of an algorithm of this novel class, which we named TH$\varepsilon$O POULA (or, simply, TheoPouLa). Finally, several experiments are presented with different types of ANNs, which show the superior performance of TheoPouLa over many popular adaptive optimization algorithms.

## 1 Introduction

Artificial neural networks (ANNs) are successfully trained when they are finely tuned via the optimization of their associated loss functions. Two aspects of such optimization tasks pose significant challenges, namely the non-convex nature of loss functions and the highly nonlinear features of many types of ANNs. Moreover, the analysis in Lovas et al. [2020] shows that the gradients of such non-convex loss functions typically grow faster than linearly and are only locally Lipschitz continuous. Naturally, stability issues are observed, which are known as the 'exploding gradient' phenomenon (Bengio et al. [1994] and Pascanu et al. [2013]), when vanilla stochastic gradient descent (SGDs) or certain types of adaptive algorithms are used for fine tuning. Section 2 provides a simple but transparent example as to why this phenomenon is observed, even when some of the most popular adaptive algorithms are employed.

One further notes that occurrences of vanishing gradients are often reported in the ANNs literature (Zhang et al. [2018] and Pascanu et al. [2013]). This phenomenon seems to particularly affect the performance of TUSLA (Lovas et al. [2020]) in our experiments when comparison is made with other popular algorithms such as AdaGrad (Duchi et al. [2011]), RMSProp (Tieleman and Hinton [2012]), ADAM (Kingma and Ba [2015]) and AMSGrad (Reddi et al. [2018]). This is observed despite TUSLA's stability properties which successfully control any potential 'exploding gradient' occurrences.

It is important to highlight that TUSLA, in contrast to the aforementioned adaptive algorithms, is built according to a new generation of (tamed) Euler approximations for

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801215 and the University of Edinburgh Data-Driven Innovation programme, part of the Edinburgh and South East Scotland City Region Deal.

stochastic differential equations (SDEs) with monotone coefficients, see Hutzenthaler et al. [2012] and Sabanis [2013], specifically targeting the class of Langevin SDEs by following the rationale of the latter article. Langevin based algorithms such as MALA (Roberts and Tweedie [1996]), ULA (Roberts and Tweedie [1996], Durmus and Moulines [2017]), which is also known as LMC (Dalalyan [2017]), TULA (Brosse et al. [2019]) and SGLD (Welling and Teh [2011]) are widely used in statistics and, in particular, in Bayesian learning. The latter algorithm has been the subject of further analysis in Raginsky et al. [2017] which highlighted the links between Langevin based algorithms and stochastic optimization in ANNs, stimulating further the development and analysis of such algorithms. For example, the incorporation of dependent data streams in the analysis of SGLD algorithms has been achieved in Barkhagen et al. [2021] and in Chau et al. [2019], and local conditions have been studied in Zhang et al. [2019] while high order schemes were developed in Sabanis and Zhang [2019] and in Li et al. [2019]. Moreover, the computational complexity of sampling algorithms as optimizers was discussed in Ma et al. [2019] within a given nonconvex setting.

Motivated by the aforementioned developments in the field, we propose a new class of adaptive algorithms which is based on Euler's polygonal approximations for Langevin SDEs. The idea of Euler's polygonal approximations for SDEs with monotone coefficients originates from the articles Krylov [1985] and Krylov [1990]. We name this new class as *polygonal unadjusted Langevin algorithms* and note that it inherits the stability properties of taming algorithms such as TUSLA. Moreover, it is versatile enough to incorporate further features to address other known shortcomings of adaptive optimizers. Mathematically, it is described as follows: Given an i.i.d. sequence of random variables $\{X_n\}_{n\geq 0}$ of interest, which typically represent available data, the algorithm follows

$$\theta_0^\lambda := \theta_0, \qquad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda G_\lambda(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \quad n \in \mathbb{N}, \tag{1}$$

where $\theta_0$ is an $\mathbb{R}^d$-valued random variable, $\lambda > 0$ denotes the step size (or learning rate) of the algorithm, $\beta > 0$ is the so-called inverse temperature, $(\xi_n)_{n\in\mathbb{N}}$ is an $\mathbb{R}^d$-valued Gaussian process with i.i.d. components and $G_\lambda : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ satisfies the following three properties:

1. There exist constants $K_1 > 0$ and $\rho_1 \geq 0$ such that $|G_\lambda(\theta, x)| \leq K_1(1+|x|)^{\rho_1}(1+|\theta|)$ for every $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$.

2. There exist constants $\gamma \geq 1/2$, $K_2 > 0$ and $\rho_2$, $\rho_3 \geq 0$ such that for all $\lambda > 0$,

$$|G_\lambda(\theta, x) - G(\theta, x)| \leq \lambda^\gamma K_2(1 + |x|)^{\rho_2}(1 + |\theta|)^{\rho_3}$$

   for every $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$, where $G$ is the (unbiased) stochastic gradient of the objective function of the optimization problem under study.

3. There exist constants $\lambda_{max}$ and $\delta \in \{1, 2\}$ such that for any $\lambda \leq \lambda_{max}$,

$$\liminf_{|\theta|\to\infty} \mathbb{E}\left[\langle \frac{\theta}{|\theta|^\delta}, G_\lambda(\theta, X_0)\rangle - \frac{2\lambda}{|\theta|^\delta}|G_\lambda(\theta, X_0)|^2\right] > 0.$$

Moreover, by considering the case where $G_\lambda(\theta, x)$ is the vector with entries $H_{\lambda,c}^{(i)}(\theta, x)$ as given by (8), for $i \in \{1, \ldots, d\}$, one obtains our new algorithm THεO POULA. Its name is formed from its description, namely Tamed Hybrid $\epsilon$-Order POlygonal Unadjusted Langevin Algorithm and its full detailed analysis (including its convergence properties) are given in Section 3.

One notes here that THεO POULA and TUSLA (Lovas et al. [2020]) satisfy the above three properties with $\delta = 2$ and $\gamma = 1/2$, whereas TULA (Brosse et al. [2019]) satisfies them with $\delta = \gamma = 1$ as it assumes only deterministic gradients (and thus the i.i.d. data sequence reduces to a constant).

THεO POULA serves as our primary example for demonstrating the stength of this new class of stochastic optimizers for ANNs. Our empirical study on various tasks with fundamental models of neural networks examines the performance of THεO POULA in comparison with the behaviour of some of the most widely used adaptive and vanilla stochastic optimizers. Our key findings can be summarised as follows (see also Sections 2 and 4 for more details):

1. **Train faster and generalize better.** THεO POULA finds its best (approximate) solution at least as fast and, in many times, faster than the other algorithms. Moreover, such a solution generalizes better than the corresponding solutions of the aforementioned optimizers.

2. **Stability.** The test error of THεO POULA remains at its lowest level, once it is achieved, during the remaining training period. In contrast, such stability is often violated by other algorithms.

3. **No vanishing or exploding gradient.** The vanishing gradient and the exploding gradient occurrences are challenging issues when training neural networks. Our experimental results show that THεO POULA does not suffer from such shortcomings when an appropriate learning rate and $\epsilon$ are chosen, and thus, there is no need for ad-hoc techniques such as gradient clipping and weight initialization.

## 2 Motivating Examples

The sparsity of gradients of neural networks is a key feature, which is extensively studied in the literature. For example, momentum methods and adaptive gradient methods such as AdaGrad (Duchi et al. [2011]), RMSProp (Tieleman and Hinton [2012]), ADAM (Kingma and Ba [2015]) and AMSGrad (Reddi et al. [2018]) have been developed to improve training speed by adequately addressing issues arising from sparse gradients. However, the local Lipschitz continuity of gradients and its effect on the performance of optimization methods are relatively under-studied. This section provides a simple, one-dimensional optimization problem that illustrates the convergence issue of these adaptive gradient methods when the gradient is locally Lipschitz continuous.

Consider the following optimization problem:

$$\min_{\theta} u(\theta) = \min_{\theta} \mathbb{E}[U(\theta, X)], \tag{2}$$

where $U : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined as

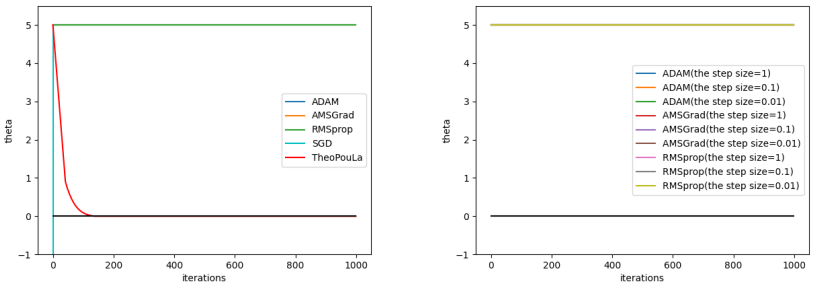
$$U(\theta, x) = \begin{cases} \theta^2 \left(1 + \mathbb{1}_{x \leq 1}\right) + \theta^{30}, & |\theta| \leq 1, \\ (2|\theta| - 1)\left(1 + \mathbb{1}_{x \leq 1}\right) + \theta^{30}, & |\theta| > 1, \end{cases}$$

and $X$ is uniformly distributed over $(-2, 2)$, that is, $f_X(x) = \frac{1}{4}\mathbb{1}_{|x| \leq 2}$. Furthermore, the stochastic gradient $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

$$G(\theta, x) = \begin{cases} 2\theta \left(1 + \mathbb{1}_{x \leq 1}\right) + 30\theta^{29}, & |\theta| \leq 1, \\ 2(1 + \mathbb{1}_{x \leq 1}) sgn(\theta) + 30\theta^{29}, & |\theta| > 1, \end{cases}$$

where $sgn(\cdot)$ is the sign function. Note that the stochastic gradient $G$ is locally Lipschitz continuous, which satisfies

$$|G(\theta, x) - G(\theta', x)| \leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'|$$




(a) default settings (b) different learning rates

Figure 1: Performance of SGD, ADAM, AMSGrad, RMSProp and THεO POULA on an artificial example with the initial value $\theta_0 = 5.0$

for all $x \in \mathbb{R}$ and $\theta, \theta' \in \mathbb{R}$. Also, the optimal value is attained at $\theta = 0$. See Appendix A for more details. Following Reddi et al. [2018], adaptive stochastic gradient methods can be generally written as follows, for $n \in \mathbb{N}$,

$$\begin{aligned} m_n &= \phi_n(G_1, \cdots, G_n), \\ V_n &= \psi_n(G_1, \cdots, G_n), \\ \theta_{n+1} &= \theta_n - \lambda_n \frac{m_n}{\varepsilon + \sqrt{V_n}} \end{aligned} \tag{3}$$

where $G_i := G(\theta_i, X_i)$ is the stochastic gradient evaluated at the $i$-th iteration, $\lambda_n$ is the step size and all operations are applied element-wise. Table 1 provides the details for some of the most popular stochastic optimization methods with corresponding averaging functions $\phi_n$ and $\psi_n$.

Table 1: Summary of stochastic optimization methods within the general framework. Note that $\widehat{v}_n = \max\{\widehat{v}_{n-1}, v_n\}$ is defined as $v_n = (1-\beta_2)v_{n-1} + \beta_2 G_n^2$.

| | SGD | RMSProp | ADAM | AMSGrad |
|---|---|---|---|---|
| $\phi_n :=$ | $G_n$ | $G_n$ | $(1-\beta_1)\sum_{i=1}^n \beta_1^{n-i} G_i$ | $(1-\beta_1)\sum_{i=1}^n \beta_1^{n-i} G_i$ |
| $\psi_n :=$ | $\mathbb{I}_n$ | $(1-\beta_2)diag(\sum_{i=1}^n \beta_2^{n-i} G_i^2)$ | $(1-\beta_2)diag(\sum_{i=1}^n \beta_2^{n-i} G_i^2)$ | $diag(\widehat{v}_n)$ |

We use SGD, ADAM, AMSGrad and RMSprop to solve the optimization problem with initial value $\theta_0 = 5$. For hyperparameters of optimization algorithms, we use their default settings provided in PyTorch. Figure 1(a) shows that SGD, ADAM, AMSGrad and RMSProp fail to converge to the optimal solution 0. Moreover, Figure 1(b) highlights that the problematic behavior cannot be resolved by adjusting the learning rate and hyperparameters within the ADAM-type framework.

Intuitively, the undesirable phenomenon occurs because, in the iterating rule (3), the denominator $\sqrt{V_n}$ dominates the numerator $m_n$, causing the vanishing gradient problem in the presence of higher-order gradients. On the contrary, SGD suffers from the exploding gradient problem. Unfortunately, it is hard to remedy the issue within the ADAM-type framework.

## 3 New Algorithm: THεO POULA

We propose a new stochastic optimization algorithm by combining ideas from taming methods specifically designed to approximate Langevin SDEs with a hybrid approach based on polygonal Euler approximations. The latter is achieved by identifying a suitable boosting function (of order $\epsilon \ll 1$) to efficiently deal with both the sparsity and the local Lipschitz continuity of (stochastic) gradients of neural networks. We proceed with the necessary preliminary information, main assumptions and formal introduction of the new algorithm.

### 3.1 Preliminaries and Assumptions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable $X$. For $1 \leq p < \infty$, $L^p$ is used to denote the usual space of $p$-integrable real-valued random variables. Fix an integer $k \geq 1$. For an $\mathbb{R}^k$-valued random variable $X$, its law on $\mathcal{B}(\mathbb{R}^k)$, i.e. the Borel sigma-algebra of $\mathbb{R}^k$, is denoted by $\mathcal{L}(X)$. Scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $|\cdot|$ standing for the corresponding norm (where the dimension of the space may vary depending on the context). For $\mu \in \mathcal{P}(\mathbb{R}^k)$ and for a non-negative measurable $f : \mathbb{R}^k \to \mathbb{R}$, the notation $\mu(f) := \int_{\mathbb{R}^k} f(\theta)\mu(\mathrm{d}\theta)$ is used. For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ denote the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures $\zeta$ on $\mathcal{B}(\mathbb{R}^{2k})$ such that its respective marginals are $\mu, \nu$. For two probability measures $\mu$ and $\nu$, the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \left( \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}, \ \mu, \nu \in \mathcal{P}(\mathbb{R}^k). \tag{4}$$

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. $\mathbb{R}^m$-valued random variables generating the filtration $(\mathcal{G}_n)_{n\in\mathbb{N}}$ and $(\xi_n)_{n\in\mathbb{N}}$ be an $\mathbb{R}^d$-valued Gaussian process with independent components. It is assumed throughout the paper that the random variable $\theta_0$, $\mathcal{G}_\infty := \sigma\left(\cup_{n\in\mathbb{N}}\mathcal{G}_n\right)$ and $(\xi_n)_{n\in\mathbb{N}}$ are independent. Let $U : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ and $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be continuously differentiable functions such that $\mathbb{E}[F(\theta, X_0] < \infty$ for any $\theta \in \mathbb{R}^d$. We consider the following optimization problem

$$\min_\theta \mathbb{E}[U(\theta, X_0)] \quad = \quad \min_\theta \left( \mathbb{E}[F(\theta, X_0)] + \frac{\eta}{2(r+1)} |\theta|^{2(r+1)} \right) \tag{5}$$

where $\theta \in \mathbb{R}^d$, $\eta \in (0, 1)$ is the regularization parameter and $r \geq \frac{q}{2} + 1$. In the context of fine tuning of ANNs, $F$ represents the loss function for the task at hand and $\theta$ denotes the vector of the ANN's parameters. Note that the regularization term, $\frac{\eta}{2(r+1)}|\theta|^{2(r+1)}$, is added in order to guarantee that the dissipativity property holds, since it is essential for the convergence analysis.[1] See Lovas et al. [2020] for more details. In particular, $r$ depends on the ANN's structure, whereas $q$ is described in Assumption 3.1. Consequently, the stochastic gradient with the regularization term is given by

$$H(\theta, x) := G(\theta, x) + \eta\theta|\theta|^{2r}$$

where $G(\theta, x) := \nabla_\theta F(\theta, x)$ for all $x \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^d$.

We proceed with our main assumptions. Our first requirement is that $G$ is locally Lipschitz continuous.

[1] One can directly assume the dissipative condition of the gradient instead of introducing the regularization term. See Raginsky et al. [2017], Xu et al. [2018] and Erdogdu et al. [2018]. However it is yet to be proven theoretically that such an assumption holds in general for ANNs.

**Assumption 3.1.** *There exists positive constant $L_1$, $\rho$ and $q \geq 1$ such that*

$$|G(\theta, x) - G(\theta', x)| \leq L_1(1 + |x|)^{\rho}(1 + |\theta| + |\theta'|)^{q-1}|\theta - \theta'|.$$

*for all $x \in \mathbb{R}^m$ and $\theta, \theta' \in \mathbb{R}^d$. Moreover, $g(\theta) := \mathbb{E}[G(\theta, X_0)]$ and $h(\theta) := \mathbb{E}[H(\theta, X_0)]$ for every $\theta \in \mathbb{R}^d$.*

Moreover, we impose conditions on the initial value $\theta_0$ and on the data process $(X_n)_{n\in\mathbb{N}}$.

**Assumption 3.2.** *The process $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_0]^{16\rho(2r+1)} < \infty$ where $\rho$ is given in Assumption 3.1. In addition, the initial condition is such that $\mathbb{E}[\theta_0]^{16(2r+1)} < \infty$.*

We refer to Appendix B for further remarks and key observations regarding the consequences of Assumptions 3.1 and 3.2. We proceed with the convergence analysis of TH$\varepsilon$O POULA by employing elements of the theory of Langevin SDEs. One first notes that, under mild conditions (satisfied by Assumptions 3.1 and 3.2), the so-called (overdamped) Langevin SDE, which is given by

$$\mathrm{d}Z_t = -h\left(Z_t\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \quad t > 0, \tag{6}$$

with a (possibly random) initial condition $\theta_0$ and with $(B_t)_{t\geq 0}$ denoting a $d$-dimensional Brownian motion, admits a unique invariant measure $\pi_\beta$. For a sufficiently large $\beta$, $\pi_\beta$ concentrates around the minimizers of (5).

### 3.2 Mechanism of TH$\varepsilon$O POULA

We introduce the mechanism of TH$\varepsilon$O POULA, which iterately updates as follows:

$$\theta_0^{\lambda} := \theta_0, \qquad \theta_{n+1}^{\lambda} := \theta_n^{\lambda} - \lambda H_{\lambda,c}(\theta_n^{\lambda}, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \quad n \in \mathbb{N}, \tag{7}$$

where $H_{\lambda,c} := (H_{\lambda,c}^{(1)}(\theta, x), \cdots, H_{\lambda,c}^{(d)}(\theta, x))^T$ is given by

$$H_{\lambda,c}^{(i)}(\theta, x) = \frac{G^{(i)}(\theta, x)}{1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|}\left(1 + \frac{\sqrt{\lambda}}{\epsilon + |G^{(i)}(\theta, x)|}\right) + \eta\frac{\theta^{(i)}|\theta|^{2r}}{1 + \sqrt{\lambda}|\theta|^{2r}}, \tag{8}$$

and $\{\xi_n\}_{n\geq 1}$ is a sequence of independent standard $d$-dimensional Gaussian random variables.

TH$\varepsilon$O POULA has three distinct features over the existing optimization methods in the literature. We give an intuitive explanation as to how these features are complementarily harmonized to improve the performance of the algorithm, and to capture two important properties of neural networks. Firstly, a taming approach is used to address stability issues due to the local Lipschitz continuity of the gradient. The new algorithm utilizes two taming functions, namely $1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|$ and $1 + \sqrt{\lambda}|\theta|^{2r}$, to control the superlinearly growing gradient and regularization term, respectively. It is worth noting that the first taming function is applied element-wise to scale the effective element-wise step size. This significantly improves the performance of our new algorithm in solving high-dimensional optimization problems such as the fine tuning of ANNs. Secondly, we have designed a suitable boosting term $\left(1 + \frac{\sqrt{\lambda}}{\varepsilon + |G^{(i)}(\theta,x)|}\right)$ to accelerate training speed. One can consider the boosting term as having a similar role to the one performed by the denominator of ADAM-like adaptive optimization methods. However, one notable difference is that our boosting term does not suppress the gradient even in the existence of the superlinearly growing gradient since $\lim_{|G|\to\infty}\left(1 + \frac{\sqrt{\lambda}}{\varepsilon + |G|}\right) = 1$, allowing us to avoid the vanishing gradient problem described in Section 2. Thirdly, a scaled Gaussian noise, $\sqrt{2\lambda\beta^{-1}}\xi_{n+1}$,

is added. This term is a consequence of the discretization of the Langevin SDE. Adding properly scaled Gaussian noise allows the new algorithm to escape local minima in a similar manner to the standard SGLD method, see Raginsky et al. [2017] and references therein. See also Neelakantan et al. [2015] for some empirical experiments.

### 3.3 Convergence Analysis

We present in this section the main convergence results of THεO POULA to $\pi_\beta$ in Wasserstein-1 and Wasserstein-2 distances as defined in (4). The convergence is guaranteed when the step size is less than $\lambda_{\max}$, which is given by

$$\lambda_{\max} = \min\left\{\frac{1}{4\eta^2}, \frac{1}{2^{14}\eta^2(_{8l}\mathcal{C}_{4l})^2}\right\}. \tag{9}$$

where ${}_n\mathcal{C}_k$ is the binomial coefficient '$n$ choose $k$' and $l = 2r+1$. Note that the step size restriction causes no issues a $\eta$ is typically very small ($\eta \ll 1$). Moreover, let $T := 1/\lambda$. Then, for any $n \in \mathbb{N}$, there exists a unique integer $m$ such as $n \in [mT, (m+1)T)$.

Theorem 3.1 and Corollary 3.1 state the non-asymptotic (upper) bounds between $\mathcal{L}\left(\theta_n^\lambda\right)$ and $\pi_\beta$. An overview of the proofs of our main results can be found in Appendix C.

**Theorem 3.1.** *Let Assumptions 3.1 and 3.2 hold. Then, there exist constants $C_1$, $C_2$, $C_3$, $\hat{c}$, $\dot{c}$ and $z_1$ such that, for every $0 < \lambda \leq \lambda_{\max}$ and $n \in \mathbb{N}$,*

$$\begin{aligned} W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \leq & \sqrt{\lambda}(z_1 + \sqrt{e^{3a}(C_1 + C_2 + C_3)}) \\ & + \hat{c}e^{-\dot{c}m}\left[1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right], \end{aligned}$$

*where $V_2$ is defined in (C.1) and $a$ is defined in Proposition B.1. The explicit form of $C_1$, $C_2$, $C_3$, $\hat{c}$, $\dot{c}$ and $z_1$ are given in Table 2.*

**Corollary 3.1.** *Let Assumptions 3.1 and 3.2 hold. Then, there exists a constant $z_2$ such that, for every $0 < \lambda \leq \lambda_{max}$ and $n \in \mathbb{N}$,*

$$\begin{aligned} W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \leq & \sqrt{e^{3a}(C_1 + C_2 + C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} \\ & + \sqrt{2\hat{c}e^{-\dot{c}m}\left(1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)}, \end{aligned}$$

*where $V_2$ is defined in (C.1). The explicit form of $z_2$ is given in Table 2.*

We are now concerned with the expected excess risk of THεO POULA generated by (7), so called the optimization error of $\theta_n^\lambda$, defined as

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \tag{10}$$

where $u(\theta) = \mathbb{E}[U(\theta, X_0)]$ and $\theta^* := \arg\min_{\theta\in\mathbb{R}^d} u(\theta)$. To derive the bound of the expected excess risk, it is again decomposed into two parts; $\mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)]$ and $\mathbb{E}[u(\theta_\infty)] - u(\theta^*)$. Here, $\theta_\infty$ follows the invariant distribution $\pi_\beta$. The following theorem describes the bound of the expected excess risk of THεO POULA.

**Theorem 3.2.** *Let Assumptions 3.1 and 3.2 hold and $\beta \geq \frac{2}{A}$. For any $n \in \mathbb{N}$, the expected excess risk of the $n$-th iterate of THεO POULA (7) is upper bounded by*

$$\begin{aligned} \mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \quad \leq \quad & \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_0|^{2l} + \frac{A_l}{\eta^2}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]\right) W_2(\mathcal{L}(\theta_n^\lambda, \pi_\beta)) \\ + \quad & \frac{1}{\beta}\left[\frac{d}{2}\log\left(\frac{eK}{A}\left(\frac{B\beta}{d} + 1\right)\right) - \log\left(1 - e^{-(R_0\sqrt{K\beta} - \sqrt{d})^2}\right)\right] \end{aligned}$$

*where $l = 2r+1$, $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$ is given in Corollary 3.1 and $a_1$, $K$, $A_l$, $R_0$ are given in Table 2.*

## 4 Emprical performance on real data sets

This section examines the performance of THεO POULA on real data sets by comparing it with those of other adaptive stochastic algorithms including ADAM, AMSGrad and RMSProp. We consider three fundamental neural network architectures with popular datasets. The first example is a convolution neural network (CNN) for CIFAR10[2] (Krizhevsky et al.), which is widely used for evaluating models or algorithms in the field of machine learning. The second example is a recurrent neural network (RNN) for language modeling on the Penn Treebank dataset[3] (Marcus et al. [1999]). The third example considers fully-connected neural networks on the auto-insurance claim data from 'freMTPL2sev" in the R package "CASdatasets"[4] to develop a predictive model of average claim sizes.

Regarding hyperparameter values of ADAM, AMSGrad and RMSProp, their default settings in Pytorch are used across experiments. More specifically, $\beta_1 = 0.9$, $\beta_2 = 0.999$ are used for ADAM and AMSGrad, and $\beta_2$ used for the RMSProp is 0.99. For THεO POULA, we set $\beta$ to be $10^{10}$. Also, $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ are considered to tune $\epsilon$.

We initially test 10 learning rates $\{10^{-1}, 5\times10^{-1}, 10^{-2}, 5\times10^{-2}, 10^{-3}, 5\times10^{-3}, 10^{-4}, 5\times 10^{-4}\}$ for the aforementioned optimization methods. If the best learning rate is determined at the end of the grid, we implement new grid points so that the best learning rate is contained in the middle of the new grid, as suggested in Wilson et al. [2017]. Moreover, we do not employ a scheme of learning rate decay as the purpose of our experiments is not to achieve state of the art results, but rather to obtain simple, transparent and interpretable results. We record the best parameter configuration which attains the minimum test loss within the fixed training budget in terms of the number of epochs. All experiments are run by a cloud computing service, which consists of Tesla V100-SXM2-16GB and Xeon Processors 2.3Ghz.

For our experiments, we consider $\eta = 0$ in (5). This is justified by the fact that some form of dissipativity may exist for specific problems such as the one considered here, although this has not been verified theoretical so far. As $\eta$ is typically extremely small, one can replace it with zero for practical implementations unless unstable paths are observed, in which case $\eta$ is switched back on and a suitable $r$ is identified. The latter task may require some effort to be completed for structurally complicated ANNs.

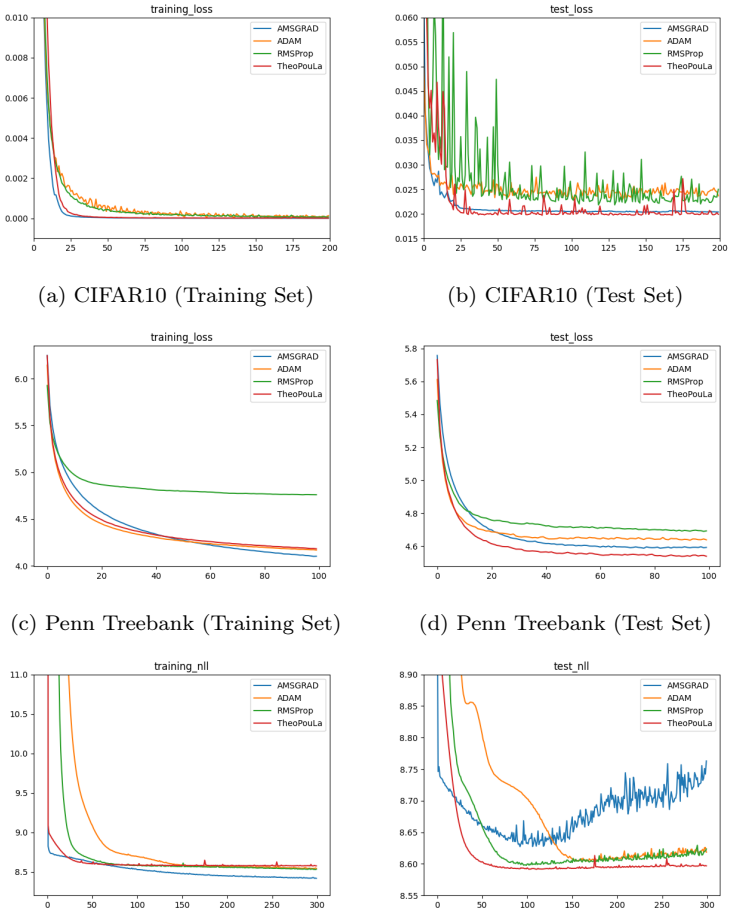
**Image classification - CIFAR10** We consider a convolution neural network, called VGG-11 (Simonya and Zisserman [2015]), which has 9 convolution layers with 5 Maxpool layers and 2 fully connected layers. Furthermore, we employ batch normalization (Ioffe and Szegedy [2015]) to prevent our models from overfitting and boost the training speed. The batch size is set to be 64.

Figure 2a and 2b show loss curves of THεO POULA, ADAM, AMSGrad and RMSProp, showing that THεO POULA trains faster and yields the lower generalization error than the other algorithms.

**Language modeling - Penn treebank** It is well known that recurrent neural networks (RNNs) suffer from problems of vanishing/exploding gradients due to their recurrent nature. This example demonstrates that the ability of the proposed algorithm to handle super-linearly growing gradients is essential for successfully training RNNs. Comparisons are made without the use of gradient clipping.

We train the Long Short Term Memory (LSTM) network with two layers and 20 time steps[5]. Each LSTM cell has 300 units and the batch size is 20. The size of embedding

[2] license: The MIT License (MIT)
[3] license: LDC User Agreement for Non-Members
[4] The dataset url is http://cas.uqam.ca/.(license: GPL-2 and GPL-3)
[5] The architecture can be found at 'https://github.com/hjc18/language_modeling_lstm'.




(a) CIFAR10 (Training Set) (b) CIFAR10 (Test Set)

(c) Penn Treebank (Training Set) (d) Penn Treebank (Test Set)

(e) Auto Insurance Claim (Training Set) (f) Auto Insurance Claim (Test Set)

Figure 2: Performance curves on training sets and test sets for three experiments.

is 300 and the dropout rate is 0.5. Figure 2d clearly show that our algorithm performs better than other algorithms in minimizing the test error. While ADAM shows fast initial progress, the loss curve become flattened after about 30 epochs. The training error of AMSGrad in Figure 2c continues to decrease, but the quality of the solution is inferior to that of our algorithm.

**Nonlinear gamma regression - auto insurance claim data** Lastly, we develop a nonlinear Gamma regression based on neural networks to predict claim severity sizes given policyholder characteristics. See Appendix E for the explanation of nonlinear gamma regression in detail. We use a fully connected feed forward neural network with 2 hidden layers and 50 neurons on each layer. Min-max normalization is used for scaling input data. The batch size is 128.

Figure 2e and 2f show the negative log likelihood (nll) curves of four different algorithms on training and test set under their best hyperparameter configurations. It is observed that THεO POULA reaches the lowest nll faster than ADAM, AMSGrad and RMSProp.

In addition, THεO POULA not only attains the lowest test negative log likelihood, but also remains at its lowest level as training progresses. One further notes that the other algorithms exhibit overfitting patterns even though their training errors continue to decrease. Our results agree with the findings of Wilson et al. [2017] where it is argued that popular adaptive optimization methods generalize worse even when these methods have lower training errors. From the above three examples, we find that, overall, THεO POULA trains faster and generalizes better than other adaptive optimization methods.

## References


M. Barkhagen, N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

H. N. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, 2011.

A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, pages 1982–2010, 2019.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. *Conference on Neural Information Processing Systems*, 2018.

M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641, 2012.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, pages 448–456, 2015.

D. Kingma and J. Ba. ADAM: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

N. V. Krylov. Extremal properties of the solutions of stochastic equations. *Theory of Probability and its Applications,*, 29(2):205–217, 1985.

N. V. Krylov. A simple proof of the existence of a solution to the Itô's equation with monotone coefficients. *Theory of Probability and its Applications,*, 35(3):583–587, 1990.

C. Li, Y. Wu, L. Mackey, and M. A. Erdogdu. Stochastic runge-kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, pages 7748–7760, 2019.

A. Lovas, I. Lytas, M. Rasonyi, and S. Sabanis. Taming neural networks with tusla: Non-convex learning via adaptive stochastic gradient langevin algorithms. *arXiv preprint arXiv:2006.14514*, 2020.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, and Ann Taylor. Treebank-3. 1999. URL https://doi.org/10.35111/gq1x-j780.

A. Neelakantan, L. Vilnis, Q.V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *Conference on Learning Theory*, 2017.

S. Reddi, S. Kale, and S. Kumar. On the convergence of ADAM and beyond. *International Conference on Learning Representations*, 2018.

G. O. Roberts and R. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

S. Sabanis. A note on tamed euler approximations. *Electronic Communications in Probability*, 18(47):1–10, 2013.

S. Sabanis and Y. Zhang. Higher order Langevin Monte Carlo algorithm. *Electronic Journal of Statistics*, 13(2):3805–3850, 2019.

K. Simonya and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. *Conference on Neural Information Processing Systems*, 2017.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Conference on Neural Information Processing Systems*, 2018.

J. Zhang, Q. Lei, and I. Dhillon. Stabilizing gradients for deep neural networks via efficient SVD parameterization. *International Conference on Machine Learning*, 2018.

Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

# Appendix

## A Details of the Experiment in Section 2

This section provides details of the experiment in Section 2. We continue to consider the optimization problem (2). One calculates that

$$u(\theta) = \begin{cases} \theta^{30} + \frac{7}{4}\theta^2, & |\theta| \leq 1, \\ \theta^{30} + \frac{7}{4}(2|\theta| - 1), & |\theta| > 1 \end{cases}$$

and

$$u'(\theta) = \begin{cases} 30\theta^{29} + \frac{7}{2}\theta, & |\theta| \leq 1, \\ 30\theta^{29} + \frac{7}{2}sgn(\theta), & |\theta| > 1. \end{cases}$$

Note that $u(\theta)$ and $u'(\theta)$ are continuous since $u(1) = \lim_{\theta \downarrow 1} u(\theta) = \frac{11}{4}$, $u(-1) = \lim_{\theta \uparrow -1} = \frac{11}{4}$, $u'(1) = \lim_{\theta \downarrow -1} u'(1) = \frac{67}{2}$ and $u'(-1) = \lim_{\theta \uparrow 1} u'(-1) = -\frac{67}{2}$. Therefore, the minimum value is attained at $\theta = 0$.

To show that $G$ is locally Lipschitz continuous, we check that for $|\theta|, |\theta'| > 1$ and $x \in \mathbb{R}^d$,

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x \leq 1})|sgn(\theta) - sgn(\theta')| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'|. \end{aligned}$$

For $|\theta|, |\theta| \leq 1$, we have

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x \leq 1})|\theta - \theta'| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'|. \end{aligned}$$

For $|\theta| \leq 1$, $|\theta| > 1$, we obtain

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x \leq 1})|\theta - sgn(\theta')| + 30|\theta^{29} - \theta'^{29}| \\ &\leq (2 + 2\mathbb{1}_{x \leq 1})|\theta - \theta'| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'| \end{aligned}$$

where the second inequality follows from the following relations

$$\begin{aligned} \theta - \theta' \leq \theta - 1 \leq 0, \quad \text{for } \theta' > 1, \\ 0 \leq \theta + 1 \leq \theta - \theta', \quad \text{for } \theta' < -1. \end{aligned}$$

## B Key Observations from Assumption 3.1 and 3.2

This section introduces some general results, that can be obtained from Assumption 3.1 and 3.2. Note that some observations can be also found in Zhang et al. [2019] and Lovas et al. [2020]. However, to make our paper self-contained, we record all the results which are necessary for the convergence analysis.

**Remark B.1.** *From Assumption 3.1, one observes that for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$*

$$|G(\theta, x)| \leq K(x)(1 + |\theta|^q)$$

*where $K(x) = 2^q(L_1(1 + |x|)^\rho + |G(0, x)|)$.*

**Remark B.2.** *From Assumptions 3.1 and 3.2, one obtains that*

$$\langle \theta,\, h(\theta) \rangle = \langle \theta,\, \mathbb{E}G(\theta, X_0) \rangle + \langle \theta,\, \eta\theta|\theta|^{2r} \rangle \geq \eta|\theta|^{2r+2} - \mathbb{E}[K(X_0)]|\theta|(1 + |\theta|^q).$$

*Furthermore, for $A = \mathbb{E}[K(X_0)]$ and $B = (3\mathbb{E}[K(X_0)])^{q+2}\eta^{-q-1}$, it holds that*

$$\langle \theta,\, h(\theta) \rangle \geq A|\theta|^2 - B. \tag{B.1}$$

**Proposition B.1.** *(Lovas et al. [2020]) Let Assumptions 3.1 and 3.2 hold. Then, for every $\theta, \theta' \in \mathbb{R}^d$,*

$$\langle \theta - \theta',\, h(\theta) - h(\theta') \rangle \geq -a|\theta - \theta'|^2$$

*where $a = L_1\mathbb{E}[(1 + |X_0|)^\rho](1 + 2|R|)^{q-1}$ and $R$ is given by*

$$R = \max\left\{ \left( \frac{2^{3(q-1)+1} L_1 \mathbb{E}[(1 + |X_0|)^\rho]}{\eta} \right)^{\frac{1}{2r-1}}, \left( \frac{2^q L_1 \mathbb{E}[(1 + |X_0|)^\rho]}{\eta} \right)^{\frac{1}{2r}} \right\}.$$

**Proposition B.2.** *(Lovas et al. [2020]) Let Assumptions 3.1 and 3.2 hold. Then, one obtains that*

$$|H(\theta, x) - H(\theta', x)| \leq L(1 + |x|)^\rho(1 + |\theta| + |\theta'|)^l|\theta - \theta'|$$

*where $L = L_1 + 8r\eta$ and $l = 2r + 1$.*

**Remark B.3.** *From Assumption 3.1 and the definition of $H$ and $H_{\lambda,c}$, one obtains that for $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$ and $i = 1, 2, \cdots, d$,*

$$\begin{aligned}
|H^{(i)}(\theta,x) - H^{(i)}_{\lambda,c}(\theta,x)| &\leq \left| G^{(i)}(\theta,x) - \frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1 + \frac{\sqrt{\lambda}}{\epsilon + |G^{(i)}(\theta,x)|}\right)\right| \\
&+ \left| \eta\theta^{(i)}|\theta|^{2r} - \eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq |G^{(i)}(\theta,x)|\frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|} + \frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{(1+\sqrt{\lambda}|G^{(i)}(\theta,x)|)(\epsilon + |G^{(i)}(\theta,x)|)} \\
&+ \eta|\theta^{(i)}||\theta|^{2r}\left|\frac{\sqrt{\lambda}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq \sqrt{\lambda}|G^{(i)}(\theta,x)|^2 + \sqrt{\lambda} + \sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{4r}
\end{aligned}$$

*which implies that*

$$\begin{aligned}
|H(\theta,x) - H_{\lambda,c}(\theta,x)|^2 &= \sum_{i=1}^d \left[\sqrt{\lambda}|G^{(i)}(\theta,x)|^2 + \sqrt{\lambda} + \sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{4r}\right]^2 \\
&\leq 3\lambda\sum_{i=1}^d \left[|G^{(i)}(\theta,x)|^4 + 1 + \eta^2|\theta^{(i)}|^2|\theta|^{8r}\right] \\
&\leq 3\lambda\left[\left(\sum_{i=1}^d |G^{(i)}(\theta,x)|^2\right)^2 + d + \eta^2|\theta|^{8r+2}\right] \\
&\leq 3\lambda\left[|G(\theta,x)|^4 + d + \eta^2|\theta|^{8r+2}\right] \\
&\leq 3\lambda\left[8|K(x)|^4(1+|\theta|^{4q}) + d + \eta^2|\theta|^{8r+2}\right]
\end{aligned}$$

**Remark B.4.** *From Assumption 3.1 and the definition of $H$ and $H_{\lambda,c}$, one calculates that for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

$$\begin{aligned}
|H(\theta,x)|^2 = |G(\theta,x) + \eta\theta|\theta|^{2r}|^2 &\leq 2|G(\theta,x)|^2 + 2\eta^2|\theta|^{4r+2} \\
&\leq 4|K(x)|^2(1+|\theta|^{2q}) + 2\eta^2|\theta|^{4r+2}
\end{aligned}$$

*and*

$$\begin{aligned}
|H_{\lambda,c}(\theta,x)|^2 &= \sum_{i=1}^d \left[\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1 + \frac{\sqrt{\lambda}}{\epsilon + |G^{(i)}(\theta,x)|}\right) + \eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right]^2 \\
&\leq \sum_{i=1}^d \left[\frac{|G^{(i)}(\theta,x)|}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|} + \frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{(1+\sqrt{\lambda}|G^{(i)}(\theta,x)|)(\epsilon + |G^{(i)}(\theta,x)|)} + \eta\frac{|\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right]^2 \\
&\leq \sum_{i=1}^d \left(|G^{(i)}(\theta,x)| + \sqrt{\lambda} + \eta|\theta^{(i)}||\theta|^{2r}\right)^2 \\
&\leq 3\sum_{i=1}^d \left(|G^{(i)}(\theta,x)|^2 + \lambda + \eta^2|\theta^{(i)}|^2|\theta|^{4r}\right) \\
&\leq 3|G(\theta,x)|^2 + 3\lambda d + 3\eta^2|\theta|^{4r+2} \\
&\leq 6|K(x)|^2(1+|\theta|^{2q}) + 3\lambda d + 3\eta^2|\theta|^{4r+2}.
\end{aligned}$$

# C Overview of the Proofs

This section provides an overview of the proofs of our main results. We begin by introducing suitable Lyapunov functions and auxiliary processes to analyze the convergence of our newly introduced algorithm. For each $m \geq 1$, define the Lyapunov function $V_m$ by

$$V_m(\theta) := (1+|\theta|^2)^{\frac{m}{2}}, \quad \theta \in \mathbb{R}^d \tag{C.1}$$

and similarly $v_m(x) = (1+x^2)^{\frac{m}{2}}$ for any real $x \geq 0$. Both functions are continuously differentiable and $\lim_{|\theta|\to\infty} \nabla V_m(\theta)/V_m(\theta) = 0$. Also, define $Z^\lambda_t = Z_{\lambda t}$, which is the time-changed Lagevin dynamics governed by

$$dZ^\lambda_t = -\lambda h(Z^\lambda_t)dt + \sqrt{2\beta^{-1}\lambda}d\tilde{B}^\lambda_t \tag{C.2}$$

where $\tilde{B}_t^\lambda = B_{\lambda t}/\sqrt{\lambda}$ is a Brownian motion.

We next define the continuous-time interpolation of the new algorithm, see (7), as

$$\mathrm{d}\bar{\theta}_t^\lambda = -\lambda H_\lambda\left(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lceil t\rceil}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{C.3}$$

with initial condition $\bar{\theta}_0^\lambda = \theta_0^\lambda$. Henceforth, $\lfloor x\rfloor$ denotes the integer part of a positive real $x$ and $\lceil x\rceil = \lfloor x\rfloor + 1$.

**Remark C.1.** *Due to the homogeneous nature of the coefficients of the continuous-time interpolation of THεO POULA (C.3) and when one selects a version of the driving Brownian motion such that it coincides with $\xi_n$ at grid points, it follows that the interpolated process (C.3) equals the process of THεO POULA (7) almost surely at grid points, i.e. $\bar{\theta}_n^\lambda = \theta_n^\lambda$ (a.s), $\forall n \in \mathbb{N}$.*

Furthermore consider the continuous-time process $\zeta_t^{s,v,\lambda}, t \geq s$ which is the solution to the SDE

$$\mathrm{d}\zeta_t^{s,v,\lambda} = -\lambda h\left(\zeta_t^{s,v,\lambda}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{C.4}$$

with initial condition $\zeta_s^{s,v,\lambda} := v, v \in \mathbb{R}^d$. Let $T := \frac{1}{\lambda}$.

**Definition C.1.** *Fix $k \in \mathbb{N}$ and define $\bar{\zeta}_t^{\lambda,k} := \zeta_t^{kT,\bar{\theta}_{kT}^\lambda,\lambda}$ where $\zeta_t^{kT,\bar{\theta}_{kT}^\lambda,\lambda}$ is defined in (C.4).*

To derive non-asymptotic (upper) bounds for $W_1\left(\mathcal{L}\left(\theta_t^\lambda\right),\pi_\beta\right)$ and $W_2\left(\mathcal{L}\left(\theta_t^\lambda\right),\pi_\beta\right)$, the following decomposition is used in terms of the auxiliary processes $\bar{\theta}_t^\lambda$, $\bar{\zeta}_t^{\lambda,n}$ and $Z_t^\lambda$ as follows:

$$W_j\left(\mathcal{L}\left(\theta_t^\lambda\right),\pi_\beta\right) \leq W_j\left(\mathcal{L}\left(\bar{\theta}_t^\lambda\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right)\right) + W_j\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^\lambda\right)\right) + W_j\left(\mathcal{L}\left(Z_t^\lambda\right),\pi_\beta\right)$$

for $j = 1, 2$.

## C.1 Primary estimates

We collect first the necessary estimates in order to obtain (upper) bounds for $W_1\left(\mathcal{L}\left(\theta_t^\lambda\right),\pi_\beta\right)$ and $W_2\left(\mathcal{L}\left(\theta_t^\lambda\right),\pi_\beta\right)$. All proofs of the lemmas in this section can be found in Appendix D. The following two lemmas provide, uniform in $n$, moment estimates of the process $(\theta_n^\lambda)_{n\geq 1}$.

**Lemma C.1.** *Let Assumptions 3.1 and 3.2 hold. Then, there exists $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0, \lambda_{\max})$ and any $n \in \mathbb{N}$,*

$$\mathbb{E}|\theta_{n+1}^\lambda|^2 \leq \left(1 - \frac{\eta}{2}\sqrt{\lambda}\right)^n \mathbb{E}|\theta_0|^2 + \left[5M_0^2 + \frac{4\sqrt{\lambda_{\max}}d}{\eta}\left(\beta^{-1} + 2 + 2\lambda_{\max}^2\right) + \frac{4(1+\lambda_{\max})\sqrt{d}M_0}{\eta} + 4\eta M_0^2\sqrt{\lambda_{\max}}\right]$$

*and, moreover,*

$$\sup_n \mathbb{E}|\theta_{n+1}^\lambda|^2 \leq \mathbb{E}|\theta_0|^2 + \left[5M_0^2 + \frac{4\sqrt{\lambda_{\max}}d}{\eta}\left(\beta^{-1} + 2 + 2\lambda_{\max}^2\right) + \frac{4(1+\lambda_{\max})\sqrt{d}M_0}{\eta} + 4\eta M_0^2\sqrt{\lambda_{\max}}\right].$$

**Lemma C.2.** *Let Assumptions 3.1 and 3.2 hold. Then, there exists $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0, \lambda_{\max})$, $n \in \mathbb{N}$, and $p \in [1, 8(2r+1)]$,*

$$\mathbb{E}|\theta_{n+1}^\lambda|^{2p} \leq (1 - \eta^2\lambda)^n \mathbb{E}|\theta_0^\lambda|^{2p} + \frac{A_p}{\eta^2}$$

*and*

$$\sup_n \mathbb{E}|\theta_{n+1}^\lambda|^{2p} \leq \mathbb{E}|\theta_0^\lambda|^{2p} + \frac{A_p}{\eta^2}$$

*where $A_p$ is given in Table 2.*

**Lemma C.3.** *Let Assumptions 3.1 and 3.2 hold. Then, there exists $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0, \lambda_{\max})$ and $n \in \mathbb{N}$,*

$$\mathbb{E}[V_4(\bar{\theta}_{nT}^\lambda)] \leq 2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2}.$$

*where $A_2$, i.e. $A_p$ for $p = 2$, is given in Table 2.*

*Proof.* From the definition of the Lyapunov function and Remark C.1, we have

$$\begin{aligned}\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{nT})] &= \mathbb{E}[(1+|\bar{\theta}^{\lambda}_{nT}|^2)^2] \\ &\leq 2+2\mathbb{E}|\bar{\theta}^{\lambda}_{nT}|^4 \\ &\leq 2+2\mathbb{E}|\theta_0|^4+2\frac{A_2}{\eta^2}\end{aligned}$$

□

Moreover, the necessary moment bounds hold also for the auxiliary process $\{\bar{\zeta}^{\lambda,n}_t\}_{t\geq nT}$.

**Lemma C.4.** *(Lemma 3.5. of Chau et al. [2019]) Let Assumptions 3.1 and 3.2 hold. Then,*

$$\begin{aligned}\mathbb{E}[V_2(\bar{\zeta}^{\lambda,n}_t)] &\leq \mathbb{E}[V_2(\theta_0)]+\frac{\tilde{c}(2)}{\bar{c}(2)}+2(C_X\eta^{-1}+2M_0^2(2+\eta)+2d(\eta\beta)^{-1}\sqrt{\lambda_{\max}})+1 \\ \mathbb{E}[V_4(\bar{\zeta}^{\lambda,n}_t)] &\leq 2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}.\end{aligned}$$

*where $\bar{c}(p)$, $\tilde{c}(p)$ are given in Table 2.*

Let $\mathcal{P}_{V_2}$ denote the subset of $\mathcal{P}(\mathbb{R}^d)$ such that every $\mu\in\mathcal{P}_{V_2}$ satisfies $\int_{\mathbb{R}^d}V_2(\theta)\mu(d\theta)<\infty$. The functional is given by

$$w_{1,2}(\mu,\nu):=\inf_{\zeta\in\mathcal{C}(\mu,\nu)}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left[1\wedge|\theta-\theta|'\right]\left[\left(1+V_2(\theta)+V_2\left(\theta'\right)\right)\zeta\left(\mathrm{d}\theta\mathrm{d}\theta'\right)\right. \tag{C.5}$$

where $\mathcal{C}(\mu,\nu)$ is defined in (4). The following lemma states the contraction property of the Langevin SDE (C.2) in $w_{1,2}$, which yields the desired result for $W_1(\mathcal{L}(Z^{\lambda}_n),\pi_\beta)$.

**Remark C.2.** *(Proposition 3.14 of Chau et al. [2019]) Let $Z'_t, t\in\mathbb{R}_+$ be the solution of the Langevin SDE (6) with initial condition $Z'_0=\theta'_0$ which is independent of $\mathcal{G}_\infty$ and $|\theta'_0|\in L^2$. Then,*

$$w_{1,2}\left(\mathcal{L}\left(Z^{\lambda}_t\right),\mathcal{L}\left(Z'_t\right)\right)\leq\hat{c}e^{-\dot{c}t}w_{1,2}\left(\mathcal{L}\left(\theta_0\right),\mathcal{L}\left(\theta'_0\right)\right)$$

*where $w_{1,2}$ is defined in (C.5).*

The following two Lemmas combined establish the required $W_1(\mathcal{L}(\bar{\theta}^{\lambda}_t),\mathcal{L}(Z^{\lambda}_t))$ estimate.

**Lemma C.5.** *Let Assumptions 3.1 and 3.2 hold. For $t>0$, there exists a unique integer $m$ such as $t\in[mT,(m+1)T)$. Then, we have for $0<\lambda<\lambda_{max}$,*

$$W_2\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_t\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_t\right)\right)\leq\sqrt{\lambda}\sqrt{e^{3a}(C_1+C_2+C_3)}$$

*where $C_1$, $C_2$, $C_3$ are given explicitly in Table 2.*

**Lemma C.6.** *Let Assumptions 3.1 and 3.2 hold. For $t>0$, there exists a unique integer $m$ such as $t\in[mT,(m+1)T)$. Then, we have for $0<\lambda<\lambda_{max}$,*

$$W_1\left(\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_t\right),\mathcal{L}\left(Z^{\lambda}_t\right)\right)\leq\sqrt{\lambda}z_1$$

*where $z_1$ is given explicitly in Table 2.*

**Lemma C.7.** *Let Assumptions 3.1 and 3.2 hold. For $t>0$, there exists a unique integer $m$ such as $t\in[mT,(m+1)T)$. Then, we have for $0<\lambda<\lambda_{max}$,*

$$W_2\left(\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_t\right),\mathcal{L}\left(Z^{\lambda}_t\right)\right)\leq\lambda^{\frac{1}{4}}z_2$$

*where $z_2$ is given explicitly in Table 2.*

## C.2 Proofs of main results

***Proof of Theorem 3.1.*** Observe that $W_1\left(\mathcal{L}\left(\theta^{\lambda}_n\right),\mathcal{L}\left(Z^{\lambda}_t\right)\right)$ is decomposed as follows:

$$W_1\left(\mathcal{L}\left(\theta^{\lambda}_n\right),\pi_\beta\right)\leq W_1\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_n\right),\mathcal{L}\left(Z^{\lambda}_n\right)\right)+W_1\left(\mathcal{L}\left(Z^{\lambda}_n\right),\pi_\beta\right).$$

Note that there exists a unique integer $m$ such that $n\in[mT,(m+1)T)$. Thus, from the results of Lemma C.5 and C.6, the first term in the right-hand side is estimated

$$\begin{aligned}W_1\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_n\right),\mathcal{L}\left(Z^{\lambda}_n\right)\right) &\leq W_1\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_n\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_n\right)\right)+W_1\left(\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_n\right),\mathcal{L}\left(Z^{\lambda}_n\right)\right) \\ &\leq W_2\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_n\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_n\right)\right)+W_1\left(\mathcal{L}\left(\bar{\zeta}^{\lambda,m}_n\right),\mathcal{L}\left(Z^{\lambda}_n\right)\right) \\ &\leq \sqrt{\lambda}(\sqrt{e^{3a}(C_1+C_2+C_3)}+z_1).\end{aligned}$$

Consequently, we derive

$$\begin{aligned} W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) &\leq \sqrt{\lambda}(\sqrt{e^{3a}(C_1+C_2+C_3)}+z_1) + w_{1,2}\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right) \\ &\leq \sqrt{\lambda}(\sqrt{e^{3a}(C_1+C_2+C_3)}+z_1) + \hat{c}e^{-\dot{c}\lambda n} w_{1,2}(\theta_0, \pi_\beta) \\ &\leq \sqrt{\lambda}(\sqrt{e^{3a}(C_1+C_2+C_3)}+z_1) + \hat{c}e^{-m\dot{c}}\left[1+\mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right] \end{aligned}$$

where Remark C.2 is used for the first inequality. □

***Proof of Corollary 3.1.*** Let $n \in [mT, (m+1)T)$. Then, Lemma C.5 and C.7 and Remark C.2 yield that

$$\begin{aligned} W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) &\leq W_2\left(\mathcal{L}\left(\bar{\theta}_n^\lambda\right), \mathcal{L}\left(Z_n^\lambda\right)\right) + W_2\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right) \\ &\leq W_2\left(\mathcal{L}\left(\bar{\theta}_n^\lambda\right), \mathcal{L}\left(\bar{\zeta}_n^{\lambda,m}\right)\right) + W_2\left(\mathcal{L}\left(\bar{\zeta}_n^{\lambda,m}\right), \mathcal{L}\left(Z_n^\lambda\right)\right) + W_2\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right) \\ &\leq \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2w_{1,2}(\mathcal{L}(Z_t^\lambda), \pi_\beta)} \\ &\leq \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{\hat{c}}e^{-\dot{c}\lambda n/2}\sqrt{2w_{1,2}(\theta_0, \pi_\beta)} \\ &\leq \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2\hat{c}e^{-\dot{c}m/2}\left(1+\mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)}. \end{aligned}$$

□

***Proof of Theorem 3.2.*** We begin by decomposing expected excess risk (10) as follows:

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \leq \mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)] + \mathbb{E}[u(\theta_\infty)] - u(\theta^*).$$

Let us focus on estimating the first part, $\mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)]$. Observe that for $\theta \in \mathbb{R}^d$

$$|\nabla u(\theta)| = |h(\theta)| \leq r_1|\theta|^{2r+1} + 2\mathbb{E}[K(X_0)]$$

by separating the cases $|\theta| \leq 1$ and $|\theta| > 1$ where $r_1 = \mathbb{E}[K(X_0)] + \eta$ due to Remark B.1. Then, we have

$$\begin{aligned} u(w) - u(v) &= \int_0^1 \langle \nabla u((1-t)v + tw),\, w - v\rangle dt \\ &\leq \int_0^1 |\nabla u((1-t)v + tw)||w - v| dt \\ &\leq \int_0^1 \left(a_1(1-t)^l|v|^l + a_1 t^l |w|^l + 2\mathbb{E}[K(X_0)]\right)|w - v| dt \\ &= \left(\frac{a_1}{l+1}|v|^l + \frac{a_1}{l+1}|w|^l + 2\mathbb{E}[K(X_0)]\right)|w - v| \end{aligned} \tag{C.6}$$

where $l = 2r+1$ and $a_1 = 2^l r_1 = 2^l(\mathbb{E}|K(X_0)| + \eta)$. Let $P$ denote the coupling between $\mu$ and $\nu$ that achieves $W_2(\mu, \nu)$ with $\mu = \mathcal{L}(\theta_n^\lambda)$ and $\nu = \mathcal{L}(\theta_\infty)$. Then, from (C.6), we obtain

$$\begin{aligned} \mathbb{E}u(\theta_n^\lambda) - \mathbb{E}u(\theta_\infty) &= \mathbb{E}_P[u(\theta_n^\lambda) - u(\theta_\infty)] \\ &\leq \mathbb{E}_P\left[\left(\frac{a_1}{l+1}|\theta_n^\lambda|^l + \frac{a_1}{l+1}|\theta_\infty|^l + 2\mathbb{E}_P[K(X_0)]\right)|\theta_n^\lambda - \theta_\infty|\right] \\ &\leq \sqrt{\mathbb{E}_P\left[\left(\frac{a_1}{l+1}|\theta_n^\lambda|^l + \frac{a_1}{l+1}|\theta_\infty|^l + 2\mathbb{E}[K(X_0)]\right)^2\right]}\sqrt{\mathbb{E}_P|\theta_n^\lambda - \theta_\infty|^2} \\ &\leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_n^\lambda|^{2l}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]\right) W_2\left(\mathcal{L}(\theta_n^\lambda), \pi_\beta\right) \\ &\leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_0|^{2l} + \frac{A_l}{\eta^2}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]\right) W_2\left(\mathcal{L}(\theta_n^\lambda), \pi_\beta\right) \end{aligned}$$

where we have used Lemma C.2 for the last inequality.

We take a similar approach in Raginsky et al. [2017] to estimate the second term. Using (B.1), we obtain

$$\langle \theta^*,\, h(\theta^*)\rangle \geq A|\theta^*|^2 - B$$

which yields

$$|\theta^*|^2 \leq \sqrt{\frac{B}{A}} \leq R_0.$$

Thus, we have

$$\begin{aligned} u(\theta^*) - u(w) &= \int_0^1 \langle \nabla u(w + t(\theta^* - w)),\, \theta^* - w\rangle dt \\ &= \int_0^1 \langle \nabla u(w + t(\theta^* - w)) - \nabla u(\theta^*),\, \theta^* - w\rangle dt \\ &= \int_0^1 \frac{1}{t-1}\langle \nabla u(w + t(\theta^* - w)) - \nabla u(\theta^*),\, w - \theta^* + t(\theta^* - w)\rangle dt. \end{aligned}$$

From Proposition B.2, we further obtain

$$\begin{aligned} -\beta(u(\theta^*) - u(w)) &= \beta|u(\theta^*) - u(w)| \\ &\le \beta \int_0^1 \frac{1}{t-1}|\langle h(w + t(\theta^* - w)) - h(\theta^*),\, w - \theta^* + t(\theta^* - w)\rangle| dt \\ &\le \beta L\mathbb{E}(1 + |X_0|)^\rho \int_0^1 (1 + |w + t(\theta^* - w)| + |\theta^*|)^l (1-t)|w - \theta^*|^2 dt \\ &\le \beta L\mathbb{E}(1 + |X_0|)^\rho \int_0^1 (1 + |w| + |\theta^* - w| + |\theta^*|)^l (1-t)|w - \theta^*|^2 dt \\ &= \beta L\mathbb{E}(1 + |X_0|)^\rho (1 + 2|\theta^*| + 2|\theta^* - w|)^l \frac{|w - \theta^*|^2}{2} \end{aligned}$$

where we have used the elementary inequality $0 \le |w| - |\theta^*| \le |\theta^* - w|$ for the last inequality.

Define $f_X$ be the density function of the multivariate normal variable $X$ with mean $\theta^*$ and covariance matrix $\frac{1}{K\beta} I_d$ where $I_d$ is the $d$-dimensional identity matrix and $K = L\mathbb{E}(1 + |X_0|)^\rho (1 + 4R_0)^l$. Then, from the above inequality, we have

$$\begin{aligned} \int_{\mathbb{R}^d} e^{\beta(u(\theta^*) - u(w))} dw &\ge \int_{\mathbb{R}^d} e^{-\beta L\mathbb{E}(1+|X_0|)^\rho (1+2|\theta^*|+2|\theta^* - w|)^l \frac{|w-\theta^*|^2}{2}} dw \\ &\ge \int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} e^{-\beta L\mathbb{E}(1+|X_0|)^\rho (1+4R_0)^l \frac{|w-\theta^*|^2}{2}} dw \\ &= \left(\frac{2\pi}{K\beta}\right)^{d/2} \int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} f_X(w) dw \\ &\ge \left(\frac{2\pi}{K\beta}\right)^{d/2} \left(1 - e^{-(R_0\sqrt{K} - \sqrt{d})^2}\right) \end{aligned}$$

where $\overline{\mathbf{B}}_r(p) = \{x \in \mathbb{R}^d | |x - p| > r\}$ and the last inequality is obtained from the following inequality thanks to the standard concentration inequality

$$\begin{aligned} P(|X - \theta^*| > R_0) &= P(\sqrt{K\beta}|X - \theta^*| > \sqrt{K\beta}R_0) \\ &\le P(\sqrt{K\beta}|X - \theta^*| - \sqrt{d} > \sqrt{K\beta}R_0 - \sqrt{d}) \\ &\le e^{-(R_0\sqrt{K\beta} - \sqrt{d})^2}. \end{aligned}$$

Define the normalizing constant $\Lambda = \int_{\mathbb{R}^d} e^{-\beta u(\theta)} d\theta$ and observe that

$$\log \Lambda \ge -\beta u(\theta^*) + \frac{d}{2}\log\left(\frac{2\pi}{K\beta}\right) + \log\left(1 - e^{-(R_0\sqrt{K\beta} - \sqrt{d})^2}\right).$$

By applying Proposition11 of Raginsky et al. [2017], we derive

$$\mathbb{E}u(\theta_\infty) - u(\theta^*) \le \frac{d}{2\beta}\log\left(\frac{eK}{A}\left(\frac{B\beta}{d} + 1\right)\right) - \frac{1}{\beta}\log\left(1 - e^{-(R_0\sqrt{K\beta} - \sqrt{d})^2}\right).$$

□

# D Proofs of Lemmas in Appendix C

***Proof of Lemma C.1.*** Define $\widehat{G}^{(i)}_{\lambda,c}(\theta, x) = \frac{G^{(i)}(\theta,x)}{1 + \sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1 + \frac{\sqrt{\lambda}}{\epsilon + |G^{(i)}(\theta,x)|}\right)$, which is part of the adaptive gradient of $H^{(i)}_{\lambda,c}(\theta, x)$. One observes that $i \in \{1, \cdots, d\}$

$$\begin{aligned} |\widehat{G}^{(i)}_{\lambda,c}(\theta, x)| &= \frac{|G^{(i)}(\theta,x)|}{1 + \sqrt{\lambda}|G^{(i)}(\theta,x)|} + \sqrt{\lambda}\frac{|G^{(i)}(\theta,x)|}{(\epsilon + |G^{(i)}(\theta,x)|)(1 + \sqrt{\lambda}|G^{(i)}(\theta,x)|)} \\ &\le \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}\frac{|G^{(i)}(\theta,x)|/\epsilon}{1 + |G^{(i)}(\theta,x)|/\epsilon} \\ &\le \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda} \end{aligned} \tag{D.1}$$

to obtain

$$\begin{aligned}\langle \theta,\, H_{\lambda,c}(\theta,x)\rangle &= \sum_{i=1}^{d}\theta^{(i)}\cdot \widehat{G}^{(i)}_{\lambda,c}(\theta,x)+\eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}}\\ &\geq \sum_{i=1}^{d}|\theta^{(i)}|\Big(-\frac{1}{\sqrt{\lambda}}-\sqrt{\lambda}\Big)+\eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}}\\ &\geq -\Big(\frac{1}{\sqrt{\lambda}}+\sqrt{\lambda}\Big)\sqrt{d}|\theta|+\eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}}\end{aligned}$$

for all $x\in\mathbb{R}^m$ and $\theta\in\mathbb{R}^d$. Then, we have

$$2\lambda\mathbb{E}\Big[\langle\frac{\theta_n^\lambda}{|\theta_n^\lambda|^2},\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle\Big|\theta_n^\lambda\Big] \geq -2\Big(\sqrt{\lambda}+\lambda^{\frac{3}{2}}\Big)\frac{\sqrt{d}}{|\theta_n^\lambda|}+2\eta\lambda\frac{|\theta_n^\lambda|^{2r}}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}}. \tag{D.2}$$

On the other hand, due to (D.1), we also get

$$\begin{aligned}|H_{\lambda,c}(\theta,x)|^2=\langle H_{\lambda,c}(\theta,x),\, H_{\lambda,c}(\theta,x)\rangle &= \sum_{i=1}^{d}\Big(\widehat{G}^{(i)}_{\lambda,c}(\theta,x)+\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\Big)^2\\ &\leq \sum_{i=1}^{d}\Big(2|\widehat{G}^{(i)}_{\lambda,c}(\theta,x)|^2+2\eta^2\frac{|\theta^{(i)}|^2|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\Big)\\ &\leq 2d\Big(\frac{1}{\sqrt{\lambda}}+\sqrt{\lambda}\Big)^2+2\eta^2\frac{|\theta|^{4r+2}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\\ &\leq 4d\Big(\frac{1}{\lambda}+\lambda\Big)+2\eta^2|\theta|^2\frac{|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}.\end{aligned} \tag{D.3}$$

which yields that

$$\begin{aligned}2\lambda\mathbb{E}\Big[-\frac{\lambda}{2|\theta_n^\lambda|^2}|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big|\theta_n^\lambda\Big] &\geq -2\lambda\Big(2d\frac{(1+\lambda^2)}{|\theta_n^\lambda|^2}+\eta^2\frac{\lambda|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\Big)\\ &\geq -4\lambda d\frac{(1+\lambda^2)}{|\theta_n^\lambda|^2}-2\lambda\eta^2.\end{aligned} \tag{D.4}$$

From (D.2) and (D.4), one calculates that

$$\begin{aligned}&2\lambda\mathbb{E}\Big[\langle\frac{\theta_n^\lambda}{|\theta_n^\lambda|^2},\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle-\frac{\lambda}{2|\theta_n^\lambda|^2}|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big|\theta_n^\lambda\Big]\\ \geq\ & -2\Big(\sqrt{\lambda}+\lambda^{\frac{3}{2}}\Big)\frac{\sqrt{d}}{|\theta_n^\lambda|}+2\eta\lambda\frac{|\theta_n^\lambda|^{2r}}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}}-4\lambda d\frac{(1+\lambda^2)}{|\theta_n^\lambda|^2}-2\lambda\eta^2=:f(\theta_n^\lambda).\end{aligned}$$

Since $f(\theta)$ tends to $2\eta\sqrt{\lambda}-2\lambda\eta^2$ as $|\theta|\to\infty$, there exists $M_0>0$ such that

$$f(\theta_n^\lambda)\geq\eta\sqrt{\lambda}-\lambda\eta^2=\eta\sqrt{\lambda}(1-\sqrt{\lambda}\eta)$$

for all $|\theta_n^\lambda|\geq M_0$ and $\lambda<\frac{1}{\eta^2}$. Moreover, for $\lambda\leq\frac{1}{4\eta^2}$, it can rewritten as there exists $M_0>0$ such that

$$f(\theta_n^\lambda)\geq\eta\sqrt{\lambda}-\lambda\eta^2=\frac{\eta\sqrt{\lambda}}{2} \tag{D.5}$$

for all $|\theta_n^\lambda|\geq M_0$.

Therefore, we have

$$\mathbb{E}\Big[\Big(2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle-\lambda^2|H_{\lambda,c}(\theta_n,X_{n+1})|^2\Big)\mathbf{1}_{|\theta_n^\lambda|\geq M_0}\Big|\theta_n^\lambda\Big]\geq\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^2,$$

implying that

$$\begin{aligned}&\mathbb{E}\Big[|\theta_{n+1}^\lambda|^2\mathbf{1}_{|\theta_n^\lambda|\geq M_0}\Big|\theta_n^\lambda\Big]\\ =\ &\mathbb{E}\Big[\Big(|\theta_n^\lambda|^2-2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n,X_{n+1})\rangle+\lambda^2|H_{\lambda,c}(\theta_n,X_{n+1})|^2+\frac{2\lambda}{\beta}|\xi_{n+1}|^2\Big)\mathbf{1}_{|\theta_n^\lambda|\geq M_0}\Big|\theta_n^\lambda\Big]\\ \leq\ &\Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)|\theta_n^\lambda|^2+\frac{2\lambda d}{\beta}\end{aligned} \tag{D.6}$$

Let us consider the case of $|\theta_n^\lambda| < M_0$. From the fact that

$$\begin{aligned}\langle \theta,\, H_{\lambda,c}(\theta,x)\rangle &= \sum_{i=1}^d \theta^{(i)}\cdot \widehat{G}_{\lambda,c}^{(i)}(\theta,x) + \eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}}\\ &\le \sum_{i=1}^d |\theta^{(i)}|\Big(\frac{1}{\sqrt{\lambda}}+\sqrt{\lambda}\Big) + \eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}}\\ &\le \Big(\frac{1}{\sqrt{\lambda}}+\sqrt{\lambda}\Big)\sqrt{d}|\theta| + \eta\frac{|\theta|^{2r+2}}{1+\sqrt{\lambda}|\theta|^{2r}} \qquad \text{(D.7)}\end{aligned}$$

and (D.3), it can be shown that

$$\begin{aligned}&\mathbb{E}\Big[|\theta_{n+1}^\lambda|^2\mathbf{1}_{|\theta_n^\lambda|<M_0}\Big|\theta_n^\lambda\Big]\\ =\;& \mathbb{E}\Big[\Big(|\theta_n^\lambda|^2 - 2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2\Big)\mathbf{1}_{|\theta_n^\lambda|<M_0}\Big|\theta_n^\lambda\Big]\\ \le\;& \Big(|\theta_n^\lambda|^2 + \frac{2\lambda d}{\beta}\Big)\mathbf{1}_{|\theta_n^\lambda|<M_0} + \mathbb{E}\Big[\Big(2\lambda|\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle| + \lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big)\mathbf{1}_{|\theta_n^\lambda|<M_0}\Big|\theta_n^\lambda\Big]\\ \le\;& |\theta_n^\lambda|^2 + \frac{2\lambda d}{\beta} + 2\Big(\sqrt{\lambda}+\lambda^{\frac{3}{2}}\Big)\sqrt{d}M_0 + 2\eta\sqrt{\lambda}M_0^2 + 4d(\lambda+\lambda^3) + 2\eta^2M_0^2\lambda\\ \le\;& \Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)|\theta_n^\lambda|^2 + \sqrt{\lambda}\Big(\frac{5\eta}{2}M_0^2 + \frac{2\sqrt{\lambda}d}{\beta} + 2(1+\lambda)\sqrt{d}M_0 + 4d(\sqrt{\lambda}+\lambda^{\frac{5}{2}}) + 2\eta^2M_0^2\sqrt{\lambda}\Big). \qquad \text{(D.8)}\end{aligned}$$

Consequently, (D.6) and (D.8) lead us to

$$\mathbb{E}\Big[|\theta_{n+1}^\lambda|^2\Big|\theta_n^\lambda\Big] \le \Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)|\theta_n^\lambda|^2 + \sqrt{\lambda}\Big(\frac{5\eta}{2}M_0^2 + \frac{2\sqrt{\lambda}d}{\beta} + 2(1+\lambda)\sqrt{d}M_0 + 4d(\sqrt{\lambda}+\lambda^{\frac{5}{2}}) + 2\eta^2M_0^2\sqrt{\lambda}\Big)$$

implying that

$$\begin{aligned}\mathbb{E}\Big[|\theta_{n+1}^\lambda|^2\Big] &\le \Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)^n\mathbb{E}|\theta_0^\lambda|^2\\ &+ \sqrt{\lambda}\Big(\frac{5\eta}{2}M_0^2 + \frac{2\sqrt{\lambda}d}{\beta} + 2(1+\lambda)\sqrt{d}M_0 + 4d(\sqrt{\lambda}+\lambda^{\frac{5}{2}}) + 2\eta^2M_0^2\sqrt{\lambda}\Big)\sum_{j=1}^\infty\Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)^j\\ &\le \Big(1-\frac{\eta\sqrt{\lambda}}{2}\Big)^n\mathbb{E}|\theta_0^\lambda|^2 + \Big(5M_0^2 + \frac{4\sqrt{\lambda}d}{\beta\eta} + 4(1+\lambda)\sqrt{d}M_0\eta^{-1} + 8d(\sqrt{\lambda}+\lambda^{\frac{5}{2}})\eta^{-1} + 4\eta M_0^2\sqrt{\lambda}\Big).\end{aligned}$$

□

***Proof of Lemma C.2.*** For any integer $p>1$, $|\theta_{n+1}^\lambda|^{2p}$ is written as

$$|\theta_{n+1}^\lambda|^{2p} = \Big(|\Delta_n|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2 + 2\langle\Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\Big)^p$$

where $\Delta_n = \theta_n^\lambda - \lambda H_{\lambda,c}(\theta_n^\lambda, X_{n+1})$. Then, we obtain

$$\begin{aligned}\mathbb{E}[|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &= \mathbb{E}\Big[\Big(|\Delta_n|^2 + \Big|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\Big|^2 + 2\langle\Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\Big)^p\Big|\theta_n^\lambda\Big]\\ &= \sum_{k_1+k_2+k_3=p}\frac{p!}{k_1!k_2!k_3!}\mathbb{E}\Big[|\Delta_n|^{2k_1}\Big|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\Big|^{2k_2}\Big(2\langle\Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\Big)^{k_3}\Big|\theta_n^\lambda\Big]\\ &\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2p\mathbb{E}\Big[|\Delta_n|^{2p-2}\langle\Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\Big|\theta_n^\lambda\Big]\\ &+ \sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\Big[|\Delta_n|^{2p-k}\Big|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\Big|^k\Big|\theta_n^\lambda\Big]\\ &\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \sum_{l=0}^{2(p-1)}\binom{2p}{l+2}\mathbb{E}\Big[\Big(|\Delta_n|^{2(p-1)-l}\Big|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\Big|^{q-1}\Big)\frac{2\lambda}{\beta}|\xi_{n+1}|^2\Big|\theta_n^\lambda\Big]\\ &= \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \binom{2p}{2}\sum_{l=0}^{2(p-1)}\binom{2(p-1)}{l}\mathbb{E}\Big[\Big(|\Delta_n|^{2(p-1)-l}\Big|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\Big|^l\Big)\frac{2\lambda}{\beta}|\xi_{n+1}|^2\Big|\theta_n^\lambda\Big]\\ &\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2^{2p-3}p(2p-1)\Big(\mathbb{E}[|\Delta_n|^{2p-2}|\theta_n^\lambda]\frac{2\lambda d}{\beta} + \Big(\frac{2\lambda}{\beta}\Big)^p\mathbb{E}|\xi_{n+1}|^{2p}\Big). \qquad \text{(D.9)}\end{aligned}$$

Define $|\Delta_n|^2 = |\theta_n^\lambda|^2 + r_n$ where $r_n = -2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2$ to write

$$\begin{aligned}\mathbb{E}\left[|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &= \sum_{k=0}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[r_n^k|\theta_n^\lambda] \\ &= |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[r_n|\theta_n^\lambda] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[r_n^k|\theta_n^\lambda]. \qquad \text{(D.10)}\end{aligned}$$

Now, focus on the case where $|\theta_n^\lambda| > M$ where

$$M := \max\{M_0, 1, \frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda}_{\max}\eta)\eta}, \frac{(1+\lambda_{\max})\sqrt{d}}{\eta(2-\eta)}, \frac{2^{2p-2}p(2p-1)d}{\eta\beta}\}$$

and $M_0$ is defined in the proof of Lemma C.1. We need the following relations to estimate the moments of $r_n$: for all $x \in \mathbb{R}^d$ and $|\theta| \geq M$,

$$\begin{aligned}\lambda^2|H_{\lambda,c}(\theta,x)|^2 &\leq 4d(\lambda+\lambda^3) + 2\eta^2\lambda|\theta|^2\frac{\lambda|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2} \\ &\leq 4d\lambda(1+\lambda^2) + 2\eta^2\lambda|\theta|^2 \\ &\leq 4d\lambda(1+\lambda^2)|\theta| + 2\eta^2\lambda|\theta|^2 \\ &\leq 2\sqrt{\lambda}\eta\left(\frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{|\theta|\eta} + \sqrt{\lambda_{\max}}\eta\right)|\theta|^2 \\ &\leq 2\sqrt{\lambda}\eta\left(\frac{2d\sqrt{\lambda_{\max}}(1+\lambda_{\max}^2)}{M\eta} + \sqrt{\lambda_{\max}}\eta\right)|\theta|^2 \\ &\leq 4\sqrt{\lambda}\eta|\theta|^2 \qquad \text{(D.11)}\end{aligned}$$

where we have used the inequality (D.3), $0 \leq \eta < 1$ and

$$M > \frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda}_{\max}\eta)\eta} \Leftrightarrow \left(\frac{2d\sqrt{\lambda_{\max}}(1+\lambda_{\max}^2)}{M\eta} + \sqrt{\lambda_{\max}}\eta\right) < 2$$

and note that $\frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda}_{\max}\eta)\eta}$ is finite due to $\lambda_{\max}$ is less than $\frac{1}{4\eta^2}$. Moreover, from (D.7), we have the following inequality

$$\begin{aligned}|2\lambda\langle\theta,\, H_{\lambda,c}(\theta,x)\rangle| &\leq 2(\sqrt{\lambda}+\lambda^{1.5})\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2\frac{\sqrt{\lambda}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}} \\ &\leq 2\sqrt{\lambda}(1+\lambda)\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2 \\ &\leq 2\sqrt{\lambda}\eta\left(\frac{(1+\lambda_{\max})\sqrt{d}}{|\theta|\eta} + \eta\right)|\theta|^2 \\ &\leq 2\sqrt{\lambda}\eta\left(\frac{(1+\lambda_{\max})\sqrt{d}}{M\eta} + \eta\right)|\theta|^2 \\ &\leq 4\sqrt{\lambda}\eta|\theta|^2 \qquad \text{(D.12)}\end{aligned}$$

where the last inequality holds since

$$M > \frac{(1+\lambda_{\max})\sqrt{d}}{\eta(2-\eta)} \Leftrightarrow \left(\frac{(1+\lambda_{\max})\sqrt{d}}{M\eta} + \eta\right) \leq 2.$$

Thus, $r_n^k$ can be written as

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|r_n|^k|\theta_n^\lambda] &= \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}\left(-2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2\right)^k\middle|\theta_n^\lambda\right] \\ &\leq \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}\left(|2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle| + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2\right)^k\middle|\theta_n^\lambda\right] \\ &\leq \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}(8\sqrt{\lambda}\eta|\theta_n^\lambda|^2)^k\middle|\theta_n^\lambda\right] \leq \lambda^{\frac{k}{2}}(8\eta)^k|\theta_n^\lambda|^{2k}\end{aligned}$$

Moreover, (D.5) implies that

$$\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}r_n|\theta_n^\lambda] \leq -\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^2,$$

equivalently,

$$p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}r_n|\theta_n^\lambda] \leq -p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p}. \qquad \text{(D.13)}$$

Using (D.13), the $L_{2p}$-norm of $\Delta_n$ conditional on $\theta_n^\lambda > M$ is given by

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &\leq |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}r_n|\theta_n^\lambda] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}|r_n|^k|\theta_n^\lambda] \\
&\leq |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\lambda^{\frac{k}{2}}(8\eta)^k|\theta_n^\lambda|^{2k} \\
&\leq |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + |\theta_n^\lambda|^{2p}\sum_{k=2}^{p}\binom{p}{k}\lambda^{\frac{k}{2}}(8\eta)^k. \qquad \text{(D.14)}
\end{aligned}$$

Choose $\lambda$ such that

$$\lambda \leq \frac{1}{(2^7\eta_p\mathcal{C}_{\lceil\frac{p}{2}\rceil})^2} = \frac{1}{2^8(8\eta)^2{}_p\mathcal{C}^2_{\lceil\frac{p}{2}\rceil}} \leq \frac{1}{2^{\frac{8}{k-1}}(8\eta)^2({}_p\mathcal{C}^2_{\lceil\frac{p}{2}\rceil})^{\frac{2}{k-1}}}$$

which is equivalent to

$$\begin{aligned}
\lambda^{\frac{k-1}{2}} &\leq \frac{1}{2^4(8\eta)^{k-1}{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}} \\
&= \frac{\eta}{2(8\eta)^k{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}}
\end{aligned}$$

for all integer $2 \leq k \leq p$. Then, since the following inequality can be obtained

$$\begin{aligned}
\sum_{k=2}^{p}{}_p\mathcal{C}_k\lambda^{\frac{k}{2}}(8\eta)^k &\leq \sum_{k=2}^{p}{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}\lambda^{\frac{k}{2}}(8\eta)^k \\
&\leq \frac{1}{2}\sum_{k=2}^{p}\sqrt{\lambda}\eta \\
&= \frac{p-2}{2}\sqrt{\lambda}\eta,
\end{aligned}$$

we have

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} \qquad \text{(D.15)}$$

and

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p-2}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2(p-2)} \leq \frac{1}{M^2}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} \qquad \text{(D.16)}$$

By combining (D.9), (D.16) and (D.15), we derive

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} \\
&+ \frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} + 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p} \\
&\leq (1-\eta\sqrt{\lambda})\left(1+\frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}\right)|\theta_n^\lambda|^{2p} \\
&+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p} \\
&\leq (1-\eta^2\lambda)|\theta_n^\lambda|^{2p} + 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p} \qquad \text{(D.17)}
\end{aligned}$$

where we used the fact that $M \geq \frac{2^{2p-2}p(2p-1)d}{\eta\beta}$ for the last inequality.

Consider the case of $|\theta_n^\lambda| \leq M$. By observing that from (D.3)

$$\mathbf{1}_{\{|\theta|\leq M\}}\lambda^2|H_{\lambda,c}(\theta,x)|^2 \leq \lambda\left(4d(1+\lambda_{\max}^2) + 2\eta^2M^2\right)$$

and

$$\begin{aligned}
\mathbf{1}_{\{|\theta|\leq M\}}|2\lambda\langle\theta,\, H_{\lambda,c}(\theta,x)\rangle| &\leq 2\lambda\sqrt{|\theta|}\sqrt{|H_{\lambda,c}(\theta,x)|} \\
&\leq 2\lambda\sqrt{M}\sqrt{|G(\theta,x)| + d\sqrt{\lambda} + 2\eta M^{2r+1}} \\
&\leq 2\lambda\sqrt{M}\sqrt{|K(x)|(1+M^q) + d\sqrt{\lambda} + 2\eta M^{2r+1}}
\end{aligned}$$

it can be shown that

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|r_n|^k\Big|\theta_n^\lambda\Big] &= \mathbb{E}\Big[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}\Big(|2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle| + \lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big)^k\Big|\theta_n^\lambda\Big]\\ &\le \mathbb{E}\Big[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}\Big(2\lambda\sqrt{M}\sqrt{K(X_{n+1})(1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1}}\\ &+ \lambda\Big(4d(1+\lambda_{\max}^2)+2\eta^2M^2\Big)\Big)^k\Big|\theta_n^\lambda\Big]\\ &\le \widetilde{D}_k\lambda^k\end{aligned}$$

where $\widetilde{D}_k = 2^{k-1}\Big((2\sqrt{M})^k(\mathbb{E}[K(X_0)](1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1})^{k/2}+(4d(1+\lambda_{\max}^2)+2\eta^2M^2)^k\Big)$. Hence, one calculates that

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|\Delta_n|^{2p}\Big|\theta_n^\lambda\Big] &\le |\theta_n^\lambda|^{2p}+\sum_{k=1}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda\le M\}}|r_n|^k|\theta_n^\lambda]\\ &\le (1-\eta^2\lambda)|\theta_n^\lambda|^{2p}+\eta^2\lambda M^{2p}+M^{2p}\lambda\sum_{k=1}^{p}\binom{p}{k}\lambda^{k-1}\widetilde{D}_k\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}_{\{\theta_n^\lambda\le M\}}|\Delta_n|^{2p-2}\Big|\theta_n^\lambda\Big] &\le \sum_{k=0}^{p-1}\binom{p-1}{k}|\theta_n^\lambda|^{2(p-1-k)}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|r_n|^k|\theta_n^\lambda]\\ &\le M^{2p-2}\sum_{k=0}^{p-1}\binom{p}{k}\widetilde{D}_k\lambda^k.\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\le (1-\eta^2\lambda)|\theta_n^\lambda|^{2p}+\eta^2\lambda M^{2p}+\lambda M^{2p}\sum_{k=1}^{p}\binom{p}{k}\lambda^{k-1}\widetilde{D}_k\\ &+ \frac{\lambda d}{\beta}2^{2p-2}p(2p-1)M^{2p-2}\sum_{k=0}^{p-1}\binom{p}{k}\lambda^k\widetilde{D}_k\\ &+ 2^{2p-3}p(2p-1)\Big(\frac{2\lambda}{\beta}\Big)^p\mathbb{E}|\xi_{n+1}|^{2p}.\end{aligned}\tag{D.18}$$

By defining

$$\begin{aligned}A_p &= \eta^2M^{2p}+M^{2p}\sum_{k=1}^{p}\binom{p}{k}\lambda_{\max}^{k-1}\widetilde{D}_k\\ &+ 2^{2p-3}p(2p-1)\Big(\frac{2dM^{2p-2}}{\beta}\sum_{k=0}^{p-1}\binom{p}{k}\lambda^k\widetilde{D}_k+\frac{2}{\beta}\Big(\frac{2\lambda_{\max}}{\beta}\Big)^{p-1}d^p(2p-1)!!\Big),\end{aligned}$$

we conclude that

$$\begin{aligned}\mathbb{E}|\theta_{n+1}^\lambda|^{2p} &\le (1-\eta^2\lambda)\mathbb{E}|\theta_n^\lambda|^{2p}+\lambda A_p\\ &\le (1-\eta^2\lambda)^n\mathbb{E}|\theta_0^\lambda|^{2p}+\lambda A_p\sum_{j=0}^{\infty}(1-\eta^2\lambda)^j\\ &\le (1-\eta^2\lambda)^n\mathbb{E}|\theta_0^\lambda|^{2p}+\frac{A_p}{\eta^2}.\end{aligned}$$

□

***Proof of Lemma C.5.*** We begin by observing that

$$\begin{aligned}
\mathbb{E}|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,m}|^2 &= -2\lambda \int_{mT}^t \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda, h(\bar{\zeta}_s^{\lambda,m}) - H_\lambda(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})\rangle ds \\
&= -2\lambda \int_{mT}^t \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda,\ h(\bar{\zeta}_s^{\lambda,m}) - h(\bar{\theta}_s^\lambda)\rangle ds \\
&- 2\lambda \int_{mT}^t \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda,\ h(\bar{\theta}_s^\lambda) - h(\bar{\theta}_{\lfloor s \rfloor}^\lambda)\rangle ds \\
&- 2\lambda \int_{mT}^t \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda,\ h(\bar{\theta}_{\lfloor s \rfloor}^\lambda) - H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})\rangle ds \\
&- 2\lambda \int_{mT}^t \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda, H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil}) - H_\lambda(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})\rangle ds \\
&\le 2\lambda a \int_{mT}^t \mathbb{E}|\bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda|^2 ds \\
&+ \frac{\lambda a}{2} \int_{mT}^t \mathbb{E}|\bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda|^2 ds + \int_{mT}^t \frac{2\lambda}{a} \mathbb{E}|h(\bar{\theta}_s^\lambda) - h(\bar{\theta}_{\lfloor s \rfloor}^\lambda)|^2 ds \\
&+ \int_{mT}^t \Big( - 2\lambda \mathbb{E}\langle \bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda,\ h(\bar{\theta}_{\lfloor s \rfloor}^\lambda) - H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})\rangle \Big) ds \\
&+ \frac{\lambda a}{2} \int_{mT}^t \mathbb{E}|\bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda|^2 ds + \int_{mT}^t \frac{2\lambda}{a} \mathbb{E}|H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil}) - H_{\lambda,c}(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})|^2 ds \\
&= 3\lambda a \int_{mT}^t \mathbb{E}|\bar{\zeta}_s^{\lambda,m} - \bar{\theta}_s^\lambda|^2 ds + \int_{mT}^t A_s^{\lambda,m} + B_s^{\lambda,m} + D_s^{\lambda,m} ds \qquad \text{(D.19)}
\end{aligned}$$

where we have used Proposition B.1 and the Young's inequality in the first inequality and

$$\begin{aligned}
A_t^{\lambda,m} &:= \frac{2\lambda}{a} \mathbb{E}|h(\bar{\theta}_t^\lambda) - h(\bar{\theta}_{\lfloor t \rfloor}^\lambda)|^2 \\
B_t^{\lambda,m} &:= -2\lambda \mathbb{E}\langle \bar{\zeta}_t^{\lambda,m} - \bar{\theta}_t^\lambda,\ h(\bar{\theta}_{\lfloor t \rfloor}^\lambda) - H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})\rangle \\
D_t^{\lambda,m} &:= \frac{2\lambda}{a} \mathbb{E}|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil}) - H_{\lambda,c}(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2.
\end{aligned}$$

In addition, from the definition of $\bar{\theta}_t^\lambda$ and (D.3), we have

$$\begin{aligned}
|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4 &\le \left( \lambda \left| \int_{\lfloor t \rfloor}^t H_{\lambda,c}(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil}) ds \right| + \sqrt{2\lambda\beta^{-1}} |\bar{B}_t^\lambda - \bar{B}_{\lfloor t \rfloor}^\lambda| \right)^4 \\
&\le 8\lambda^2 \left( \lambda^2 |H_{\lambda,c}(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})|^4 + 4\beta^{-2} |\bar{B}_t^\lambda - \bar{B}_{\lfloor t \rfloor}^\lambda|^4 \right) \\
&\le 8\lambda^2 \left( (4d(1+\lambda^2) + 2\eta^2 |\bar{\theta}_{\lfloor s \rfloor}^\lambda|^2)^2 + 4\beta^{-2} |\bar{B}_t^\lambda - \bar{B}_{\lfloor t \rfloor}^\lambda|^4 \right) \\
&\le \lambda^2 2^5 \left( 2^3 d^2 (1+\lambda^4) + \eta^4 |\bar{\theta}_{\lfloor s \rfloor}^\lambda|^4 + \beta^{-2} |\bar{B}_t^\lambda - \bar{B}_{\lfloor t \rfloor}^\lambda|^4 \right)
\end{aligned}$$

which yields

$$\sqrt{\mathbb{E}|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} \le \widetilde{C}_1 \lambda \qquad \text{(D.20)}$$

where $\widetilde{C}_1 = 2^{5/2} \sqrt{8d^2(1+\lambda_{\max}^4) + \eta^4(\mathbb{E}|\theta_0|^4 + A_2/\eta^2) + \frac{3}{\beta^2} d^2}$.

Using Proposition B.2, $A_t^{\lambda,m}$ can be bounded as follows:

$$\begin{aligned}
A_t^{\lambda,m} &\le \frac{2\lambda}{a} L_X \mathbb{E}[(1 + |\bar{\theta}_t^\lambda| + |\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{2l} |\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^2] \\
&\le \frac{2\lambda}{a} L_X \sqrt{\mathbb{E}(1 + |\bar{\theta}_t^\lambda| + |\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{4l}} \sqrt{\mathbb{E}|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} \\
&\le \frac{2\lambda}{a} L_X 3^{2l} \sqrt{(1 + \mathbb{E}|\bar{\theta}_t^\lambda|^{4l} + \mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4l})} \sqrt{\mathbb{E}|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} \\
&\le C_1 \lambda^2 \qquad \text{(D.21)}
\end{aligned}$$

where $L_X = L^2 2^{2\rho-1}(1 + \mathbb{E}|X_0|^{2\rho})$ and $C_1 = \frac{2}{a} L_X 9^l \sqrt{(1 + 2\mathbb{E}|\bar{\theta}_0^\lambda|^{4l} + 2\frac{A_{2l}}{\eta^2})} \widetilde{C}_1$ and (D.20) is used for the last inequality.

To estimate $B_t^{\lambda,m}$, we observe that

$$\begin{aligned}
B_t^{\lambda,m} &= -2\lambda\mathbb{E}\langle\bar{\zeta}_t^{\lambda,m} - \bar{\theta}^\lambda_{\lfloor t\rfloor},\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle \\
&\quad - 2\lambda\mathbb{E}\langle\bar{\theta}^\lambda_{\lfloor t\rfloor} - \bar{\theta}^\lambda_s,\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle \\
&\le -2\lambda\mathbb{E}\Big[\mathbb{E}\Big[\langle\bar{\zeta}_t^{\lambda,m} - \bar{\theta}^\lambda_{\lfloor t\rfloor},\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle\Big|\bar{\zeta}_t^{\lambda,m}, \bar{\theta}^\lambda_{\lfloor t\rfloor}\Big]\Big] \\
&\quad - 2\lambda\mathbb{E}\Big[\langle\bar{\theta}^\lambda_{\lfloor t\rfloor} - \bar{\theta}^\lambda_t,\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle\Big] \\
&\le -2\lambda\mathbb{E}\Big[\langle\lambda\int_{\lfloor t\rfloor}^t H_\lambda(\bar{\theta}^\lambda_{\lfloor s\rfloor}, X_{\lceil s\rceil})ds - \sqrt{\frac{2\lambda}{\beta}}\tilde{B}^\lambda_{t-\lfloor t\rfloor},\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle\Big] \\
&\le -2\lambda^2\mathbb{E}\Big[\langle H_\lambda(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil}),\, h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})\rangle\Big] \\
&\le 2\lambda^2\sqrt{\mathbb{E}|H_\lambda(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})|^2}\sqrt{\mathbb{E}|h(\bar{\theta}^\lambda_{\lfloor t\rfloor}) - H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})|^2} \\
&\le 2\lambda^2\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{2q}) + 3\lambda d + 3\eta^2\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{4r+2}}\sqrt{4\mathbb{E}|H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})|^2} \\
&\le 4\lambda^2\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{2q}) + 3\lambda d + 3\eta^2\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{4r+2}} \\
&\times \sqrt{4\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}^\lambda_{\lfloor s\rfloor}|^{2q}) + 2\eta^2\mathbb{E}|\bar{\theta}^\lambda_{\lfloor s\rfloor}|^{4r+2}} \\
&\le C_2\lambda^2
\end{aligned} \tag{D.22}$$

where

$$\begin{aligned}
C_2 &= 4\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\theta_0|^{2q} + \frac{A_q}{\eta^2}) + 3\lambda d + 3\eta^2|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{4r+2}} \\
&\times \sqrt{4\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\theta_0|^{2q} + \frac{A_q}{\eta^2}) + 2\eta^2\Big(\mathbb{E}|\bar{\theta}^\lambda_0|^{4r+2} + \frac{A_{2r+1}}{\eta^2}\Big)}.
\end{aligned}$$

Note that we have used the independence of $\bar{\theta}^\lambda_{\lfloor s\rfloor}$ and $X_{\lceil s\rceil}$ to obtain the second inequality, and used Remark B.4 and Lemma C.2 to calculate the bound of $\mathbb{E}|H_\lambda(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})|^2$ and $\mathbb{E}|H(\bar{\theta}^\lambda_{\lfloor t\rfloor}, X_{\lceil t\rceil})|^2$.

Moreover, $D_t^{\lambda,m}$ can be estimated as follows, from Remark B.3,

$$\begin{aligned}
D_t^{\lambda,m} &\le \frac{6\lambda^2}{a}\Big[8\mathbb{E}|K(X_0)|^4(1+\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{4q}) + d + \eta^2\mathbb{E}|\bar{\theta}^\lambda_{\lfloor t\rfloor}|^{8r+2}\Big] \\
&\le C_3\lambda^2
\end{aligned} \tag{D.23}$$

where the independence of $\bar{\theta}^\lambda_{\lfloor s\rfloor}$ and $X_{\lceil s\rceil}$ is used and $C_3$ is given by

$$C_3 = \frac{6}{a}\Big[8\mathbb{E}|K(X_0)|^4(1+\mathbb{E}|\bar{\theta}^\lambda_0|^{4q} + A_{2q}/\eta^2) + d + \eta^2(\mathbb{E}|\bar{\theta}^\lambda_0|^{8r+2} + A_{4r+1}/\eta^2)\Big].$$

Plugging (D.21), (D.22) and (D.23) into (D.19), one can derive

$$\begin{aligned}
\mathbb{E}|\bar{\theta}^\lambda_t - \bar{\zeta}_t^{\lambda,m}|^2 &\le 3\lambda a\int_{mT}^t \mathbb{E}|\bar{\theta}^\lambda_s - \bar{\zeta}_s^{\lambda,m}|^2 ds + \int_{nT}^t (C_1+C_2+C_3)\lambda^2 ds \\
&\le 3\lambda a\int_{mT}^t \mathbb{E}|\bar{\theta}^\lambda_s - \bar{\zeta}_s^{\lambda,m}|^2 ds + (C_1+C_2+C_3)\lambda < \infty
\end{aligned}$$

where the second inequality follows from the fact that $(t-mT) \le T = \frac{1}{\lambda}$ and the use of Grownwall's inequality gives

$$\mathbb{E}|\bar{\theta}^\lambda_t - \bar{\zeta}_t^{\lambda,m}|^2 \le c\lambda$$

where $c = e^{3a}(C_1+C_2+C_3)$. □

***Proof of Lemma C.6.*** Since $Z_t^\lambda = \bar{\zeta}_t^{\lambda,0}$ and $t \in [mT, (m+1)T)$, we can write

$$\begin{aligned}
W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,m}\right), \mathcal{L}\left(Z_t^\lambda\right)\right) &\le \sum_{k=1}^m W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) \\
&\le \sum_{k=1}^m w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right)
\end{aligned}$$

where we have used the fact $W_1(\mu,\nu) \le w_{1,2}(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}_{V_2}$ for the second inequality. Using Remark C.2 and $\lambda(t-kT) \ge m-k$, we further have

$$\begin{aligned} w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) &\le \hat{c}e^{-\dot{c}\lambda(t-kT)} w_{1,2}\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\ &\le \hat{c}e^{-\dot{c}(m-k)} w_{1,2}\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\ &\le \hat{c}e^{-\dot{c}(m-k)} W_2\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \sqrt{\mathbb{E}\left|1+V_2(\bar{\theta}_{kT}^{\lambda})+V_2(\bar{\zeta}_{kT}^{\lambda,k-1})\right|^2} \\ &\le \hat{c}e^{-\dot{c}(m-k)} W_2\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\ &\times \left[1+\sqrt{\mathbb{E}[V_4(\bar{\theta}_{kT}^{\lambda})]}+\sqrt{\mathbb{E}[V_4(\bar{\zeta}_{kT}^{\lambda,k-1})]}\right] \\ &\le \hat{c}e^{-\dot{c}(m-k)}\sqrt{\lambda}\sqrt{e^{3a}(C_1+C_2+C_3)}\left[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right. \\ &+ \left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right] \end{aligned} \tag{D.24}$$

where the Cauchy-Schwarz inequality is applied to the third inequality and Lemma C.3, C.5 and C.4 are used for the last inequality. By combining the two inequalities above, we obtain

$$\begin{aligned} W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,m}\right), \mathcal{L}\left(Z_t^{\lambda}\right)\right) &\le \hat{c}\sqrt{\lambda}\sqrt{e^{3a}(C_1+C_2+C_3)}\left[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right. \\ &+ \left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right]\sum_{k=1}^{m} e^{-\dot{c}(m-k)} \\ &\le z_1\sqrt{\lambda} \end{aligned}$$

where $z_1 = \frac{\hat{c}}{1-\exp(-\dot{c})}\sqrt{e^{3a}(C_1+C_2+C_3)}\left[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right]$. □

*Proof of Lemma C.7.* We begin by observing that

$$\begin{aligned} W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) &\le \sqrt{2w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right)} \\ &\le \lambda^{1/4}e^{-\dot{c}(m-k)/2}\left[\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right.\right. \\ &+ \left.\left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2} \end{aligned}$$

where we have used the fact $W_2 \le \sqrt{2w_{1,2}}$ for the first inequality, and the second inequality follows from (D.24). Consequently, we derive

$$\begin{aligned} W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,m}\right), \mathcal{L}\left(Z_t^{\lambda}\right)\right) &\le \sum_{k=1}^{m} W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) \\ &\le \lambda^{1/4}\left[\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right.\right. \\ &+ \left.\left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2}\sum_{k=1}^{m} e^{-\dot{c}(m-k)/2} \\ &\le \lambda^{1/4} z_2 \end{aligned}$$

where

$$z_2 = \frac{\sqrt{\hat{c}}e^{3a/4}(C_1+C_2+C_3)^{1/4}}{1-\exp(-\dot{c}/2)}\left[\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2}.$$

□

# E Details of Experiments in Section 4

Gamma regression is an important tool for modeling a strictly positive target variable. In particular, it is widely used in insurance business to predict insurance claim sizes. Suppose that we construct a

dataset from $m$ independent observations:

$$\{\mathbf{z}_i\}_{i=1}^m := \{\mathbf{x}_i, y_i\}_{i=1}^m$$

where $x_i \in \mathbb{R}^p$ is the feature set of covariates, and $y_i$ is the average claim size. Note that we use only data with positive severity from the "CASdatasets".

Let $Y_i$ be the severity for a policy holder $i$. It is assumed that the mean function of $Y_i$ is written in terms of a regression function $S(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ with the log link:

$$\ln \mu(\mathbf{x}_i) := \ln \mathbb{E}[Y_i|\mathbf{x}_i] = S(\mathbf{x}_i)$$

or equivalently,

$$\mu_i = \exp\{S(\mathbf{x}_i)\}.$$

Since we are interested in a nonlinear regression, no linearity condition is imposed on $S(\mathbf{x}_i)$. Instead, $S(\cdot)$ is expressed by neural networks to incorporate non-linearity in explanatory variables, which is estimated by minimizing the negative log-likelihood:

$$\min_{\mathbf{w}\in\Theta, \phi\in\mathbb{R}^+} \sum_{i=1}^m \ln f_Y(y_i; \mu_i, \phi)$$

where $\mathbf{w}$ is the parameter of the neural network, $\Theta \in \mathbb{R}^d$ is the feasible set of parameters for the neural network and $f(y; \mu, \phi)$ is the gamma distribution with mean $\mu$ and dispersion parameter $\phi$:

$$f_Y(y; \mu, \phi) = \frac{1}{y\Gamma(\phi^{-1})} \left(\frac{y}{\mu\phi}\right)^{\frac{1}{\phi}} e^{-\frac{y}{\mu\phi}}.$$

# F Table of Constants

Table 2 displays full expressions for constants which appear in the main results of this paper. In addition, Table 3 shows all main constants and their dependencies on key parameters such as $d$, $\beta$, the moments of $K(X_0)$ and $\eta$.

Table 2: Explicit expression for constants with $\hat{c}$ and $\dot{c}$ from Proposition 3.14 of Chau et al. [2019].

| Symbol | Full Expression |
|---|---|
| $M$ | $\max\left\{M_0, 1, \frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda}_{\max}\eta)\eta}, \frac{(1+\lambda_{\max})\sqrt{d}}{\eta(2-\eta)}, \frac{2^{2p-2}p(2p-1)d}{\eta\beta}\right\}$ |
| $\tilde{D}_k$ | $2^{k-1}\Big[(2\lambda_{\max}\sqrt{M})^k(\mathbb{E}[K(X_0)](1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1})^{k/2}$ $+(4d(1+\lambda_{\max}^2)+2\eta^2M^2)^k\Big], \quad k=1,\cdots,8(2r+1)$ |
| $A_p$ | $\eta^2M^{2p}+M^{2p}\sum_{k=1}^{p}\begin{pmatrix}p\\k\end{pmatrix}\lambda_{\max}^{k-1}\tilde{D}_k$ $+2^{2p-3}p(2p-1)\Big(\frac{2dM^{2p-2}}{\beta}\sum_{k=0}^{p-1}\begin{pmatrix}p\\k\end{pmatrix}\lambda_{\max}^k\tilde{D}_k+\frac{2}{\beta}\Big(\frac{2\lambda_{\max}}{\beta}\Big)^{p-1}d^p(2p-1)!!\Big),$ for $p=1,\cdots,8(2r+1)$ |
| $\overline{M}_p$ | $\sqrt{\frac{1}{3}+4B/(3A)+4d/(3A\beta)+4(p-2)/(3A\beta)}$ |
| $\bar{c}(p)$ | $\frac{Ap}{4}, \quad p=1,\cdots,8(2r+1)$ |
| $\tilde{c}(p)$ | $\frac{3}{4}Apv_p(\overline{M}_p), \quad p=1,\cdots,8(2r+1)$ |
| $C_1$ | $\frac{L^2 2^{2\rho+5/2}3^{2l}}{a}(1+\mathbb{E}\lvert X_0\rvert^{2\rho})\sqrt{(1+2\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4l}+2\frac{A_{2l}}{\eta^2})}$ $\times\sqrt{8d^2(1+\lambda_{\max}^4)+\eta^4(\mathbb{E}\lvert\theta_0\rvert^4+A_2/\eta^2)+\frac{3}{\beta^2}d^2}$ |
| $C_2$ | $4\sqrt{6\mathbb{E}\lvert K(X_0)\rvert^2(1+\mathbb{E}\lvert\theta_0\rvert^{2q}+\frac{A_q}{\eta^2})+3\lambda_{\max}d+3\eta^2\lvert\bar{\theta}_{\lfloor t\rfloor}^{\lambda}\rvert^{4r+2}}$ $\times\sqrt{4\mathbb{E}\lvert K(X_0)\rvert^2(1+\mathbb{E}\lvert\theta_0\rvert^{2q}+\frac{A_q}{\eta^2})+2\eta^2\Big(\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4r+2}+\frac{A_{2r+1}}{\eta^2}\Big)}$ |
| $C_3$ | $\frac{6}{a}\Big[8\mathbb{E}\lvert K(X_0)\rvert^4(1+\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4q}+A_{2q}/\eta^2)+d+\eta^2(\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{8r+2}+A_{4r+1}/\eta^2)\Big].$ |
| $z_1$ | $\frac{\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}}{1-\exp(-\hat{c})}\Big[1+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Big]$ |
| $z_2$ | $\frac{\sqrt{\dot{c}}e^{3a/4}(C_1+C_2+C_3)^{1/4}}{1-\exp(-\dot{c}/2)}\Big(1+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Big)^{\frac{1}{2}}$ |
| $a_1$ | $2^l(\mathbb{E}[K(X_0)]+\eta)$ |
| $R_0$ | $\inf\{y\geq\sqrt{B/A}: y^2(1+4y)^l>\frac{d+1}{\beta L\mathbb{E}(1+\lvert X_0\rvert)^{\rho}}\}$ |
| $K$ | $L\mathbb{E}[(1+\lvert X_0\rvert)^{\rho}](1+4R_0)^l$ |

Table 3: Main constants and their dependency to key parameters

| Constant | Key parameters | | | |
|---|---|---|---|---|
| | $d$ | $\beta$ | Moments of $X_0$ | $\eta$ |
| $A$ | - | - | $\mathcal{O}(\mathbb{E}K(X_0))$ | - |
| $B$ | - | - | $\mathcal{O}(\mathbb{E}K(X_0)^{q+2})$ | $\mathcal{O}(\frac{1}{\eta^{q+1}})$ |
| $R$ | - | - | $\mathcal{O}(\mathbb{E}\lvert X_0\rvert^{\rho})$ | $\mathcal{O}(\frac{1}{\eta^{2r-q}})$ |
| $a$ | - | - | $\mathcal{O}\left(\mathbb{E}\lvert X_0\rvert^{\rho(q-1)}\right)$ | $\mathcal{O}(\frac{1}{\eta^{(2r-q)(q-1)}})$ |
| $A_p$ | $poly(d)$ | $\mathcal{O}(\frac{d}{\beta})$ | $\mathcal{O}(\mathbb{E}K\lvert X_0\rvert^{p/2})$ | $\mathcal{O}(\frac{1}{\eta^{p-2}})$ |
| $\dot{c}$ | $\mathcal{O}(e^{-d})$ | Inherited from contraction estimates in Eberle et al. [2019] | | |
| $\hat{c}$ | $\mathcal{O}(e^{d})$ | Inherited from contraction estimates in Eberle et al. [2019] | | |